TH Köln

University of Applied Sciences

Fakultät für Wirtschafts- und Rechtswissenschaften

Masterarbeit / Thesis (Drei-Monats-Arbeit)

zur Erlangung
des akademischen Grades Master of Science
im Studiengang Versicherungswesen

„Statistische Analyse ausgewählter sozioökonomischer Daten in Bezug auf Krankenhausaufenthalte“

Erstprüfer	Professor Dr. Jan-Philipp Schmidt
Zweitprüfer	Professor Dr. Jürgen Strobel
vorgelegt am	31. August 2018
von	Lars Dirking
aus	[redacted]
Matrikel-Nr.	[redacted]
Telefon-Nr.	[redacted]
E-Mail-Adresse	[redacted]

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abb.	Abbildung
AUC	Area under the curve
BMI	Body-Mass-Index
CRAN	Comprehensive R Archive Network
DIW	Deutsches Institut für Wirtschaftsforschung
FN	false negatives
FP	false positives
GLM	Generalized Linear Model
KH	Krankenhaus
LinReg	Lineare Regression
LM	Linear Model
LogReg	Logistische Regression
OECD	Organisation für wirtschaftliche Zusammenarbeit und Entwicklung
PKV	Private Krankenversicherung
ROC-Kurve	Receiver-Operating-Characteristic-Kurve
Sens	Sensitivität
SOEP	Sozio-oekonomisches Panel
Spec	Spezifität
TN	true negatives
TP	true positives
VN	Versicherungsnehmer
VU	Versicherungsunternehmen

1. Einleitung

Deutschlandweit gibt es 1.951 Krankenhäuser (Stand: 2016). Insgesamt lagen die Bruttokosten der Krankenhäuser im Jahr 2016 bei rund 101,66 Mrd. Euro.[1] Dabei machten stationäre Leistungen den größten Anteil aus. Dem statistischen Bundesamt zufolge betrugen die sogenannten bereinigten Kosten, welche die Bruttokosten abzüglich nichtstationärer Kosten sind, etwa 87,84 Mrd. Euro. Bestimmte Kosten, beispielsweise für Ambulanz, wissenschaftliche Forschung und Lehre werden somit von den Kosten für stationäre Leistungen abgegrenzt.[2]

Diese immensen Kosten sind in den letzten Jahren kontinuierlich angestiegen. So betrugen die bereinigten Kosten im Jahr 1996 noch 48,36 Mrd. Euro.[3] Im Jahr 2016 gab es rund 19,5 Mio. Behandlungsfälle in deutschen Krankenhäusern. Daraus ergeben sich bereinigte Kosten pro Fall in Höhe von etwa 4.497 Euro.

Es gibt zahlreiche Ursachen, warum Menschen ein Krankenhaus aufsuchen müssen. Häufig wird ein Krankenhaus im Falle eines akuten Notfalls, beispielsweise eines Unfalls oder eines Herzinfarktes, oder für Behandlungen, die nicht ambulant durchgeführt werden können, aufgesucht. Überdies werden auch Vorsorgeuntersuchungen in Krankenhäusern durchgeführt. Einige Krankheiten sind zu einem gewissen Anteil genetisch bedingt, wie zum Beispiel die Vererbbarkeit des Brustkrebs-Risikos.[4] Andere Krankheiten, wie etwa Adipositas oder Hypertonie, werden auch durch das Verhalten des einzelnen Menschen beeinflusst. So lässt sich das Risiko für diese Erkrankungen durch eine gesunde Ernährung und ausreichende sportliche Aktivität deutlich verringern.[5]

Die Kosten für Krankenhausaufenthalte werden zu großen Teilen von der gesetzlichen (GKV) und privaten Krankenversicherung (PKV) übernommen. Deshalb haben auch diese Institutionen ein Interesse daran, zu erfahren, inwiefern es Verbindungen zwischen dem Verhalten oder anderen sozioökonomischen Daten eines Versicherungsnehmers (VN) und den jeweiligen Krankenhauskosten gibt. Ziel dieser Arbeit ist es deshalb, das Zusammenspiel von bestimmten sozioökonomischen Merkmalen und Krankenhausaufenthalten zu ermitteln. Hierzu werden Daten des Sozio-oekonomischen Panels (SOEP) des Deutschen Instituts für Wirtschaftsforschung (DIW) verwendet, um statistische Zusammenhänge zu analysieren.

---

[1] Vgl. Statistisches Bundesamt (2017), S. 12
[2] Vgl. Statistisches Bundesamt (2017a)
[3] Vgl. Statistisches Bundesamt (2017), S. 9
[4] Vgl. Krebsinformationsdienst (2013)
[5] Vgl. Robert Koch-Institut (2015), S. 206

Zunächst wird anhand von Sekundärliteratur gezeigt, welche gesundheitsbeeinflussenden Faktoren es in Deutschland gibt und was häufige Ursachen für Krankheit bzw. Krankenhausaufenthalte sind. Im Zuge dessen wird außerdem kurz auf die Tarifmerkmale der privaten Krankenversicherung sowie auf deren Einfluss auf die Prämienkalkulation eingegangen.

Im darauf folgenden Kapitel wird der Datensatz des DIW Berlin analysiert. Dabei werden einerseits einzelne Merkmale und deren Entwicklung im Laufe der Jahre sowie die Verteilung ausgewählter Merkmale auf bestimmte Personengruppen beobachtet. Andererseits werden auch Korrelationen zwischen bestimmten Merkmalen und Hospitalisierungswahrscheinlichkeiten untersucht. Dadurch soll ein Versicherungsunternehmen (VU) den Bestand der Versicherten besser gliedern und im Idealfall sogar Aussagen zur Wahrscheinlichkeit eines bevorstehenden Krankenhausaufenthaltes treffen können.

Hierzu werden zunächst die mathematischen Grundlagen zweier Klassifikationsmethoden, der linearen und der logistischen Regressionsanalyse, erläutert. Daraufhin werden diese Methoden mithilfe des Programms R auf zuvor selektierte sozioökonomische Merkmale des vorliegenden Datensatzes angewendet, um Hospitalisierungswahrscheinlichkeiten zu ermitteln. Dabei wird sowohl ein Basismodell als auch ein Alternativmodell aufgestellt, um eine Vergleichbarkeit zu schaffen.

Diese beiden Modelle beinhalten teilweise verschiedene Merkmale, die auf Basis der Ergebnisse der Sekundärforschung dazu geeignet scheinen, einen Einfluss auf die Möglichkeit eines Krankenhausaufenthaltes im Folgejahr zu haben. Die Ergebnisse der verschiedenen Modelle werden anhand von Gütemaßen überprüft, um zu ermitteln, welches Modell am besten zur Prädiktion von Krankenhauswahrscheinlichkeiten geeignet ist.

In einem abschließendem Kapitel wird unter anderem untersucht, ob es einem Versicherer möglich ist, solch eine statistische Analyse durchzuführen und welche Herausforderungen und Chancen dies mit sich bringt. Dies beinhaltet neben Fragen des Datenschutzes und der Datenerhebung auch Präventionsmaßnahmen im Falle einer hohen Hospitalisierungswahrscheinlichkeit. Darüber hinaus wird eine Schlussbetrachtung der erzielten Forschungsergebnisse vorgenommen und ein kurzer Ausblick auf künftige Entwicklungen bei der Verwendung sozioökonomischer Daten zur Prädiktion von Krankenhausaufenthalten gegeben.

2. Gesundheitsbeeinflussende Faktoren

Es gibt zahlreiche Faktoren, die auf den Gesundheitszustand des Menschen einwirken. Diese können einerseits erblich bedingt sein, andererseits beispielsweise vom Verhalten des Menschen oder seines Umfelds abhängig sein. Im Folgenden wird zum einen die derzeitige Situation sowie die Entwicklung der letzten Jahre dargelegt, zum anderen werden die wichtigsten Tarifmerkmale der privaten Krankenversicherung und deren Relevanz für die Prämienkalkulation erläutert.

2.1. Status quo in Deutschland

Häufig auftretende Krankheiten und deren Ursachen unterscheiden sich von Staat zu Staat. Dabei ist neben geografischen und klimatischen Bedingungen vor allem relevant, wie weit die medizinische Versorgung und das Gesundheitssystem des jeweiligen Staates fortgeschritten ist. Deutschland ist eine Industrienation und hat eine vergleichsweise gute medizinische Versorgung sowie ein funktionierendes System aus gesetzlicher und privater Krankenversicherung.

Doch auch in Deutschland bleibt es nicht aus, dass Krankheiten auftreten. Betroffene suchen bei leichteren Krankheiten nicht sofort ein Krankenhaus auf, sondern genesen zuhause oder lassen sich ambulant in einer Arztpraxis behandeln. Da die vorliegende Arbeit sozioökonomische Daten in Bezug auf Krankenhausaufenthalte analysiert, wird der Fokus auf stationäre Leistungen und jene ambulante Leistungen, die im Krankenhaus durchgeführt werden, gesetzt.

In Deutschland gab es im Jahr 2016 rund 19,5 Mio. Behandlungsfälle in Krankenhäusern.[6] Da dies mit hohen Kosten, vor allem für Versicherer, einhergeht, werden im Folgenden häufige Krankheiten der in Deutschland lebenden Bevölkerung dargelegt und darüber hinaus untersucht, was die Ursachen für bestimmte Erkrankungen sind. Als Ursachen werden hierbei weniger direkte Ursachen wie beispielsweise eine Infektion herangezogen, sondern vielmehr indirekte, sozioökonomische Ursachen wie das Verhalten oder die Krankheitsgeschichte der Person.

Die folgende Tabelle zeigt die Anzahl der häufigsten Hauptdiagnosen in Krankenhäusern:

[6] Vgl. Statistisches Bundesamt (2017a)

Diagnose	Anzahl
Herzinsuffizienz	455.680
Psychische und Verhaltens-störungen durch Alkohol	322.608
Vorhof-flimmern und Vorhof-flattern	304.755
Intrakranielle Verletzung	282.678
Hirninfarkt	258.480
Pneumonie, Erreger nicht näher bezeichnet	243.430
Angina pectoris	242.490
Sonstige chronische obstruktive Lungen-krankheit	238.552
Cholelithiasis	234.546
Rückenschmerzen	225.634
Essentielle (primäre) Hypertonie	225.393
Akuter Myokar-dinfarkt	219.157
Atherosklerose	200.998
Chronische ischämische Herz-krankheit	196.194
Bösartige Neubildung der Bronchien und der Lunge	193.697

Tabelle 1: Häufigste Hauptdiagnosen in Krankenhäusern 2016[7]

Die Anzahl der Krankenhausaufenthalte aufgrund von Geburten bleibt in dieser Tabelle unbeachtet. Man erkennt, dass in Deutschland bei zahlreichen stationär behandelten Patienten Erkrankungen des Herz-Kreislauf-Systems diagnostiziert wurden. Darüber hinaus machen alkoholbedingte Krankenhausaufenthalte sowie Lungenerkrankungen einen großen Anteil aus. Außerdem sind die für Deutschland typischen Erkrankungen des Skeletts sowie Bluthochdruck sehr häufige Diagnosen.

Ein wesentliches Merkmal, das die Gesundheit eines Menschen beeinflusst, ist der sozioökonomische Status oder Sozialstatus des Einzelnen. Der Sozialstatus beschreibt die Stellung eines Menschen innerhalb einer Gesellschaft. Er wird häufig durch sein Bildungsniveau, seinen Beruf und die Einkommenssituation bestimmt.[8] Ähnlich wie die Ungleichverteilung von Einkommen, so sind auch Gesundheitschancen und Krankheitsrisiken nicht auf die Gesamtbevölkerung gleichverteilt. In der Regel sind Menschen mit niedrigerem Sozialstatus auch gesundheitlich schlechter gestellt und haben überdies eine geringere Lebenserwartung als Menschen mit höherem Status.[9]

Die folgende Abbildung (Abb.) zeigt, wie stark der soziale Status mit dem Gesundheitszustand zusammenhängt:

[7] Statistisches Bundesamt (2017b)
[8] Lampert, Thomas et al. (2013), S.632
[9] Vgl. Robert Koch-Institut (2015), S. 149

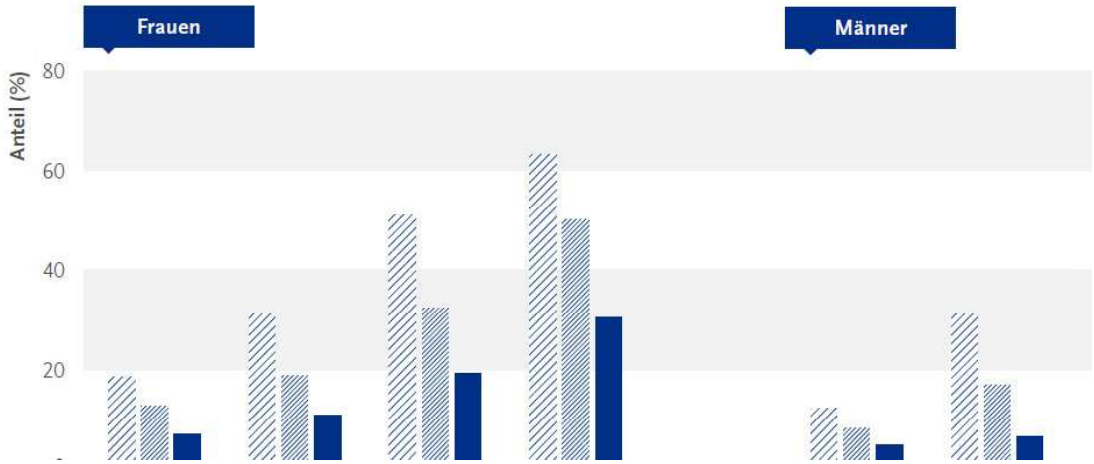

Abbildung 1: Häufigkeit eines mittleren bis schlechten Gesundheitszustandes[10]

Man erkennt die deutliche Tendenz der Menschen mit niedrigem sozioökonomischem Status, die eigene Gesundheit als schlecht einzustufen. Tatsächlich ist der Gesundheitszustand dieser Menschen schlechter, sie leiden beispielsweise häufiger unter chronischen Krankheiten, Diabetes mellitus, degenerativen Erkrankungen sowie psychischen Erkrankungen.[11]

Gründe für diese Differenzen liegen auch häufig im Gesundheitsverhalten. So verfügen Menschen der unteren Sozialgruppen tendenziell über weniger gesundheitsbezogenes Wissen, haben andere Einstellungen und haben weniger Mittel zur Verfügung, um etwa gesünder einzukaufen. Außerdem nehmen sie seltener Präventionsangebote wahr.[12] Darüber hinaus üben sie oftmals Berufe aus, in denen die körperliche Belastung hoch ist. Dies führt unter anderem auch dazu, dass Menschen mit niedrigem Sozialstatus sportliche Aktivitäten seltener ausüben. All diese Verhaltensweisen tragen auch dazu bei, dass die Verbreitung von Adipositas bei dieser Gruppe besonders ausgeprägt ist.[13]

Sowohl mit dem sozioökonomischen Status als auch mit dem Gesundheitszustand ist die berufliche Tätigkeit, die ein Mensch ausübt, verknüpft. Durch eine Erwerbstätigkeit wird eine geregelte Zeitstruktur ermöglicht, außerdem steigt das soziale Ansehen. Aufgrund des vorhandenen Einkommens kann überdies eine gesündere Lebensweise sowie eine bessere Absicherung der Gesundheit gewährleistet werden. Darüber hinaus hat auch der Arbeitgeber ein Interesse, dass der Arbeitnehmer gesund und somit arbeitsfähig ist.

---

[10] Robert Koch-Institut (2015), S. 150
[11] Vgl. Robert Koch-Institut (2015), S. 150
[12] Vgl. Robert Koch-Institut (2015), S. 151
[13] Vgl. Mensink, G.B.M. et al. (2013), S. 792

Ein weiterer Indikator für die Gesundheit ist deshalb die Anzahl der Tage, an denen der Arbeitnehmer oder auch ein Arbeitsloser arbeitsunfähig war, obgleich dies auch nicht-gesundheitliche Gründe haben kann.[14] Es kann zum einen beobachtet werden, dass Personen mit einem höheren Schul- und Ausbildungsabschluss pro Jahr tendenziell weniger Tage arbeitsunfähig sind als jene mit niedrigerem Abschluss (s. Anhang, Abb. 15 und 16). Zum anderen fällt auf, dass Arbeitslose deutlich häufiger arbeitsunfähig sind als Arbeitnehmer.[15]

Arbeitslose haben außerdem ein erhöhtes Risiko, sowohl psychische als auch körperliche Erkrankungen zu entwickeln. Dabei kann eine Erkrankung sowohl die Ursache als auch die Folge der Arbeitslosigkeit sein.[16] Eine Arbeitslosigkeit verschlechtert häufig den sozioökonomischen Status. Da es vielen Arbeitslosen, insbesondere Langzeitarbeitslosen, an finanziellen Mitteln mangelt, weisen einige einen gesundheitsriskanteren Lebensstil auf, um Geld zu sparen. Auf der anderen Seite ist die Quote der Raucher in der Gruppe der Arbeitslosen deutlich höher, was einerseits mitunter hohe Ausgaben nach sich zieht, andererseits die Gesundheit zusätzlich stark schädigt.[17]

Ein weiterer gesundheitsbeeinflussender Faktor ist die Herkunft eines Menschen. Derzeit leben in Deutschland rund 22,5% Menschen mit Migrationshintergrund.[18] Generell lässt sich sagen, dass die Zusammensetzung dieser Personengruppe sehr heterogen ist. Entsprechend gibt es auch Unterschiede, aus welchem Land die Menschen mit Migrationshintergrund tatsächlich stammen. Sie sind im Durchschnitt jünger als die deutsche Bevölkerung, was das Krankheitsrisiko grundsätzlich verringert. Nicht zuletzt aufgrund der sprachlichen Barriere, die zu Beginn besonders hoch ist, stehen Menschen mit Migrationshintergrund jedoch häufig vor Herausforderungen, was sich auch auf den Sozialstatus und somit den Gesundheitszustand auswirkt.

Das Gesundheitsverhalten ist häufig ein anderes als das der einheimischen Bevölkerung. So kommt es auf das Herkunftsland und die entsprechende individuelle Person an, ob das Verhalten gesundheitsförderlich oder –schädigend ist. Gesundheitsleistungen werden von ihnen tendenziell seltener in Anspruch genommen als von Menschen ohne Migrationshintergrund, was unter anderem auf die genannte

[14] Vgl. Robert Koch-Institut (2015), S. 158
[15] Vgl. Techniker Krankenkasse (2017), S. 36f.
[16] Vgl. Robert Koch-Institut (2003), S. 6ff.
[17] Vgl. Kroll, L.E. / Lampert, T. (2012), S. 5
[18] Vgl. Bundeszentrale für politische Bildung (2018), S.6

Sprachbarriere, aber ebenfalls auf Diskriminierungserfahrungen zurückzuführen ist.[19]

Ein weiterer Indikator, welcher Einfluss auf die Gesundheit haben kann, stellt die Wohnsituation dar. Wesentliche Faktoren sind hier der Straßenverkehr, dessen Lärmbelästigung und Luftschadstoffe gesundheitsschädigend sind, sowie die Wohnung selbst, bei der unter anderem Größe, Lüftung und sanitäre Anlagen von Relevanz sind.[20] Lärm, insbesondere, wenn er über viele Jahre hinweg schlafmindernd wirkt, kann sowohl psychische als auch physische Schäden nach sich ziehen.

Ist die Wohnung bzw. das Haus unzureichend oder falsch belüftet, so kann es vermehrt zu schädlichen Schimmelpilzen kommen. Diese können zu Erkrankungen der Atemwege sowie zu Allergien des Betroffenen führen.[21] Betrachtet man die sozialen Unterschiede, so fällt auf, dass die Wohnsituation bei Menschen von niedrigerem sozioökonomischem Status oft schlechter ist und infolgedessen die dadurch bedingten Erkrankungen häufiger auftreten.[22]

Neben den genannten Einflüssen spielt der Body-Mass-Index (BMI) eine wichtige Rolle bei Frage nach der Wahrscheinlichkeit bzw. Häufigkeit bestimmter Erkrankungen. Personen, die keiner bis wenig sportlicher Aktivität nachgehen, weisen weitaus häufiger Übergewicht auf als diejenigen, die ein angemessenes Maß an Sport ausüben. Außerdem kann beobachtet werden, dass umso weniger Sport betrieben wird, je niedriger der Sozialstatus der Person ist. Dies liegt unter anderem daran, dass Menschen dieser Gruppe häufiger körperlichen Arbeiten nachgehen, weshalb nach Feierabend seltener Sport betrieben wird.

Auch die Ernährung hat einen großen Einfluss auf das Risiko von Übergewicht sowie den Gesundheitszustand im Allgemeinen. Zwar sind Lebensmittel in Deutschland relativ preiswert, dennoch kauft längst nicht jeder Konsument gesund ein. Besonders Fleisch wird in zu großen Mengen verzehrt, vor allem von Männern.[23] Dementgegen steht ein zu geringer Konsum von frischem Gemüse, Obst und Milchprodukten.

Zu den größten Faktoren, die auf den Gesundheitszustand einwirken, zählen das Rauchen sowie der Genuss von Alkohol. Rauchen, auch das Passivrauchen, stellt

[19] Vgl. Robert Koch-Institut (2015), S. 181
[20] Vgl. Robert Koch-Institut (2015), S. 185
[21] Vgl. Robert Koch-Institut (2007), S. 3
[22] Vgl. Bolte, G. / Kohlhuber, M. (2008), S. 5ff.
[23] Vgl. Robert Koch-Institut (2015), S. 195

das bedeutendste einzelne Gesundheitsrisiko dar und ist in Deutschland der Hauptgrund für vorzeitige Sterblichkeit.[24] Im Jahr 2013 starben durch Tabak und dessen Folgen rund 121.000 Menschen.[25] Der Konsum von Tabak kann insbesondere Erkrankungen des Herz-Kreislauf-Systems, der Atemwege sowie Krebserkrankungen hervorrufen. In Deutschland raucht rund 28,3% der Bevölkerung.[26] Dabei rauchen junge Erwachsene rauchen häufiger als über 65-jährige.

Ebenso stellt der übermäßige Konsum von Alkohol ein massives Risiko für Krankheiten dar. Er erhöht, je nach Intensivität, die Wahrscheinlichkeit von Entzündungen der Bauchspeicheldrüse und der Magenschleimhaut und kann Hirnschäden sowie bestimmte Krebsarten hervorrufen. Darüber hinaus wird während des Rausches das Risiko für Unfälle, Verletzungen und gewalttätige Auseinandersetzungen stark erhöht.

In Deutschland sterben jährlich rund 74.000 Menschen an den Folgen von Alkohol, wobei jedoch drei Viertel der Fälle eine Kombination von Alkohol und Tabak darstellen.[27] Zur Feststellung, ob jemand übermäßig viel Alkohol zu sich nimmt, stellt der Pro-Kopf-Konsum einen wichtigen Indikator dar. Es ist jedoch zu beachten, dass Befragte bei Selbstangaben dazu tendieren, ihre Angaben aufgrund sozialer Erwünschtheit zu untertreiben.

Besonders auffällig ist, dass die Quoten beim Konsum sowohl von Tabak als auch von Alkohol bei Menschen mit niedrigem sozioökonomischem Status weitaus höher sind als bei jenen mit höherem Sozialstatus. Gründe hierfür sind einerseits ein entsprechendes Umfeld, andererseits ein zu geringes Gesundheitswissen bezüglich der Folgen des Konsums. Insgesamt lässt sich sagen, dass es viele Faktoren gibt, die auf den Gesundheitszustand und damit letztlich ebenfalls auf die Sterblichkeit eines Menschen, einwirken.

Die Kosten für stationäre Leistungen sind in Deutschland im internationalen Vergleich grundsätzlich relativ gering.[28] Dennoch werden in Deutschland vergleichsweise viele Mittel für Krankenhausaufenthalte aufgewendet. Dies liegt daran, dass die Anzahl der Fälle sehr hoch ist, Deutschland hat also eine starke Aktivität in Krankenhäusern. Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) hat für das Jahr 2012 ermittelt, dass es in Deutschland rund 240 Kranken-

[24] Vgl. Robert Koch-Institut (2015), S. 218
[25] Vgl. Deutsches Krebsforschungszentrum (2015), S. 1
[26] Vgl. Kotz, D. / Böckmann, M. / Kastaun, S. (2018), S. 1
[27] Vgl. Deutsche Hauptstelle für Suchtfragen e.V. (2018), S. 1
[28] Vgl. Technische Universität Berlin (2014), S. 14f.

hausentlassungen pro 1000 Einwohner gab. Dies lag rund 50 Prozent über dem Durchschnitt des OECD-Raumes.[29]

Bezieht man den eingangs beschriebenen Wert von rund 19,5 Mio. Behandlungsfällen im Jahr 2016 auf die Gesamtbevölkerung [30] dieses Jahres, so stellt man fest, dass es pro 1000 Einwohner etwa 237 Behandlungsfälle gab. Demnach ist die Anzahl von Behandlungsfällen und somit auch Versicherungsfällen nach wie vor hoch. Dies wiederum wirkt sich indirekt auf die Leistungsausgaben von PKV und GKV aus. So betrugen im Jahr 2016 im Fall der GKV die Ausgaben für Krankenhausbehandlungen insgesamt rund 72,95 Mrd. Euro[31]. Im Fall der PKV lagen die Aufwendungen für stationäre Leistungen bei 7,59 Mrd. Euro[32]. Hierbei sind Leistungen wie etwa Krankenhaustagegeld nicht einbezogen.

Die Kosten eines Krankenhauses, welche zur Bestimmung der Leistungshöhe der Versicherer maßgeblich sind, setzen sich hauptsächlich aus Personal- und Sachkosten zusammen. Darüber hinaus existieren kleinere Positionen wie Zinsen und ähnliche Aufwendungen sowie Steuern.[33] Da im Bereich der privaten Krankenversicherung teilweise andere, über die Leistungen der gesetzlichen Krankenversicherung hinausgehende, Leistungen erstattet werden, sind die jeweiligen Kosten unterschiedlich auf die beiden Versicherungssysteme verteilt.

[29] Vgl. Kumar, A. / Schoenstein, M. (2013), S. 9
[30] Vgl. Statistisches Bundesamt (2018)
[31] Vgl. Bundesgesundheitsministerium (2018), S. 1
[32] Vgl. Verband der Privaten Krankenversicherung (2017), S. 53
[33] Vgl. Statistisches Bundesamt (2017), S. 3

2.2. Tarifmerkmale der privaten Krankenversicherung

Offensichtlich ist der Gesundheitszustand einer Person von vielen Faktoren abhängig. Eines der grundlegenden Prinzipien von Versicherung ist, dass Versicherungsnehmer ihr individuelles Risiko, welches einen mitunter stark volatilen Verlauf hat, durch eine fixe Prämie substituieren. Somit wird ein drohender Ruin des VN verhindert. Es findet ein Ausgleich im Kollektiv statt, da die Einzelrisiken der versicherten Personen zu einem gewissen Grad voneinander unabhängig sind. Aufgrund dessen sinkt das Risiko mit steigender Kollektivgröße, sodass insgesamt weniger Kapital zur Deckung benötigt wird.[34]

Der Gesundheitszustand einer versicherten Person hat Auswirkungen auf den Erwartungswert und die Standardabweichung des Schadenverlaufs. Informationen zur derzeitigen Gesundheit sowie vorangegangene Ereignisse wie Krankenhausaufenthalte sind demnach notwendig, um den erwarteten Jahresschaden dieser versicherten Person zu kalkulieren. Sie gehören zu den wichtigsten Risikomerkmalen der PKV. Diese Risikomerkmale werden in der Krankenversicherung auch Tarifmerkmale genannt.[35] Das Versicherungsunternehmen hat hierüber jedoch zunächst keine Informationen.

Im Rahmen des Antragsprozesses werden deshalb zahlreiche Informationen über den VN abgefragt. Aufgrund der Versicherungspflicht ist es einem Krankenversicherer nicht möglich, einen Antragssteller mit einer schlechten Risikolage abzulehnen, zumindest nicht, wenn er den Basistarif beantragt. Damit das Kollektiv jedoch nicht unter dem höheren Risiko dieses VN leidet, werden für bestimmte Krankheiten in der Regel Risikozuschläge vereinbart. Diese Zuschläge können entweder als Prozentsatz der Prämie oder aber als absoluter Betrag vereinbart werden.

Wichtige Risikomerkmale sind jedoch nicht nur Vorerkrankungen und der derzeitige Gesundheitszustand einer Person. Wie im vorangegangenen Kapitel erläutert, spielen zahlreiche andere Faktoren eine Rolle bei der Bestimmung, wie häufig jemand ins Krankenhaus kommt bzw. welche Kosten für das Versicherungsunternehmen oder die Krankenkasse dadurch entstehen können.

Wesentliche Tarifmerkmale, die in ihrer Gesamtheit letztlich zur Prämienhöhe des Versicherungsnehmers beitragen, sind vor allem auch die versicherten Leistungen. In der privaten Krankenversicherung haben im Gegensatz zur gesetzlichen Absiche-

[34] Vgl. Becker (2017), S. 27
[35] Vgl. Becker (2017), S. 30

rung längst nicht alle Versicherungsnehmer die gleichen Leistungen versichert. So kann ein VN unter anderem wählen, welchen Leistungsbereich er absichern möchte. Dadurch kann er beispielsweise festlegen, ob er ambulante oder stationäre Behandlungen versichern möchte. Außerdem wirken Wartezeiten, Höchstsätze und Selbstbehalte sowie die konkret abgedeckten Leistungen auf die Prämienhöhe ein.

Ein weiteres Tarifmerkmal stellt das Alter der Person dar. Zur Selektierung können Versicherte auch in Altersgruppen zusammengefasst werden. Der Gesundheitsberichterstattung des Bundes zufolge, lagen im Jahr 2015 die jährlichen Krankheitskosten eines 30- bis 45-jährigen Menschen lediglich bei rund 2.240 Euro, während die eines 65- bis 85-jährigen bei rund 8.350 Euro lagen (s. Anhang, Abb. 17). Deshalb gibt es bei der kapitalgedeckten privaten Krankenversicherung Alterungsrückstellungen, welche dazu beitragen, dass die Versicherungsprämie mit steigendem Alter nicht zu stark ansteigt.

Obwohl seit dem 21.12.2012 durch das Allgemeine Gleichbehandlungsgesetz (AGG) eine Ungleichbehandlung von Mann und Frau verboten wird,[36] spielt dieses Tarifmerkmal dennoch eine Rolle und fließt bei der Ermittlung der Risikoprämie in die Kalkulation ein. Damit Männer und Frauen jedoch keine unterschiedlichen Prämien zahlen müssen, wird dieses Tarifmerkmal bei Ermittlung der Individualprämie des VN wieder auf das andere Geschlecht umgelegt, sodass letztlich für den einzelnen Versicherungsnehmer diesbezüglich Gleichheit besteht.

Das individuelle Risiko des Antragsstellers wird unter anderem auch durch seinen Beruf definiert. Stark körperliche Berufe oder Berufe mit hohem psychischem Druck tragen beispielsweise dazu bei, dass Schadenerwartungswert und somit die Prämie des Versicherungsnehmers steigt. Um dieses Risiko zu kompensieren, wird deshalb für bestimmte Berufe ein entsprechender Risikozuschlag vereinbart.

Neben den genannten, objektiven Risikofaktoren, gibt es darüber hinaus auch subjektive Risikomerkmale. Dabei handelt es sich um jene Merkmale, die stark vom individuellen Verhalten des Versicherungsnehmers abhängig sind. Diese Merkmale sind zwar risikowirksam, jedoch jederzeit einseitig durch den Versicherungsnehmer änderbar oder durch den Versicherer nicht sicher feststellbar. Beispiele hierfür sind unter anderem das Leistungs-Inanspruchnahmeverhalten des VN sowie seine Lebensweise im Allgemeinen. [37]

[36] Vgl. §§ 19-20 AGG
[37] Vgl. Milbrodt, H. / Kniep, T. (2005), S. 40f.

3. Analyse der sozioökonomischen Daten

Im folgenden Kapitel werden ausgewählte sozioökonomische Daten mithilfe statistischer Methoden analysiert. Bei den Daten handelt es sich um einen Teil der Langversion des Deutschen Sozio-oekonomischen Panels (SOEPlong[38]) des Deutschen Instituts für Wirtschaftsforschung. Das im Jahr 1925 gegründete Institut wird überwiegend aus öffentlichen Mitteln finanziert und erforscht wirtschafts- und sozialwissenschaftliche Zusammenhänge in gesellschaftlich relevanten Themenfeldern.[39]

Beim Sozio-oekonomischen Panel handelt es sich um eine Wiederholungsbefragung, bei der deutsche Haushalte seit 1984 jährlich zu Einkommen, Erwerbstätigkeit, Bildung und Gesundheit befragt werden. Es werden zahlreiche, in ausgewählten Haushalten lebende, Menschen befragt. Um langfristige Trends erkennen zu können, werden nach Möglichkeit jedes Jahr dieselben Personen befragt.[40]

Aufgrund der großen Anzahl an Befragten und des langen Zeitraums, sind einige der Personen bereits aus dem Pool der Befragten ausgeschieden und neue hinzugekommen. Durchgeführt wird die Datenerhebung von Kantar Public Deutschland, welches in Deutschland das Nachfolgeinstitut von TNS Infratest Sozialforschung und TNS Infratest Politikforschung ist. Auch Kantar Public gibt an, ein unabhängiges Institut zu sein.[41]

3.1. Untersuchte Merkmale

Der Datensatz umfasst aktuell 2.915 Merkmale, von denen jedoch nicht alle jedes Jahr erfragt wurden. Darüber hinaus gibt es teilweise Überschneidungen von Merkmalen, etwa, wenn sich die Frage im Laufe der Jahre leicht geändert hat. Außerdem gibt es Fragen, die sich auf Themengebiete oder Sachverhalte beziehen, welche zu früheren Befragungszeitpunkten noch nicht bestanden haben, beispielsweise Fragen zum Mindestlohn.[42] Andererseits sind zahlreiche frühere Fragen heute weniger relevant als sie es vor rund 30 Jahren noch waren. Dies betrifft unter anderem Fragen bezüglich der Unterschiede zwischen der Bundesrepublik und der Deutschen Demokratischen Republik.

---

[38] Vgl. Schupp et al. (2017)
[39] Vgl. Deutsches Institut für Wirtschaftsforschung (a)
[40] Vgl. Deutsches Institut für Wirtschaftsforschung (b)
[41] Vgl. Kantar Public
[42] Vgl. Deutsches Institut für Wirtschaftsforschung (2015)

Aufgrund der immens hohen Anzahl an Merkmalen, wurden für diese Arbeit nur jene selektiert, die für Forschungsfrage der Arbeit als relevant erachtet wurden. Folglich werden 35 Merkmale statistisch analysiert. Dabei handelt es sich neben Merkmalen wie der Hospitalisierung oder der Anzahl der Krankheitstage im Vorjahr insbesondere um Daten zum sozioökonomischen Verhalten.

Die Merkmale wurden auf Basis der Ergebnisse der Sekundärforschung der vorangegangenen Kapitel festgelegt. Zu den sozioökonomischen Daten gehören daher beispielsweise das letzte Monatsgehalt, die Regelmäßigkeit von Sport, das Ernährungs- und Raucherverhalten einer Person sowie wie die Einschätzung des eigenen Gesundheitszustandes. Darüber hinaus werden verhaltensunabhängige Merkmale wie das Geschlecht, das Alter oder die Größe einer Person in die Analyse einbezogen.

Die Merkmale des Datensatzes sind jeweils mit Abkürzungen betitelt. Dessen vollständige Bedeutungen werden über die Internetseite der Survey-Gruppe SOEP geliefert.[43] So hat beispielsweise das Kürzel „ple0007“ die Bedeutung „Körpergewicht in kg“. Zur besseren Übersichtlichkeit wurden alle relevanten Merkmale entsprechend umbenannt, sodass die Namen der Kürzel in den folgenden Ausführungen nicht von Relevanz sind.

Merkmale werden in quantitative und qualitative Merkmale unterschieden. Bei qualitativen Merkmalen sind die Merkmalsausprägungen jeweils einzelne Kategorien. Jeder Merkmalsträger wird wiederum genau einer Kategorie zugeordnet. Kann man die Kategorien eines qualitativen Merkmals hierarchisch anordnen, so handelt es sich um ein ordinalskaliertes Merkmal. Ist es jedoch nicht möglich, die Merkmalsausprägungen in eine Rangfolge zu bringen, etwa im Falle des Geschlechts, werden die Merkmale als nominalskaliert bezeichnet.[44]

Anders als bei qualitativen Merkmalen, nehmen die Ausprägungen bei quantitativen bzw. metrischen Merkmalen reelle Zahlen an. Es kann demnach nicht nur eine Rangbildung erfolgen, sondern auch eine Unterscheidung nach der Höhe der Werte. Es existiert folglich ein Abstandsmaß, was im Rahmen eines Vergleichs mehrerer Merkmalsträger konkretere Aussagen ermöglicht. Es sind verschiedene Rechenoperatoren möglich, zum Beispiel die Ermittlung eines Mittelwertes.

---

[43] Vgl. https://paneldata.org/soep-long/data/pl
[44] Vgl. Handl, A. / Kuhlenkasper, T. (2017), S. 17

Die kategorischen Merkmalsausprägungen ordinalskalierter Merkmale werden häufig als ganze Zahlen dargestellt bzw. können in solche umgeformt werden. Es wird hierbei beispielsweise die Skala „sehr gut“ bis „sehr schlecht“ zu einer Skala mit den Ausprägungen „1“ bis „5“ konvertiert. Dies ermöglicht das Rechnen mit diesen Größen, obgleich zu erwähnen ist, dass der Abstand zwischen eins und zwei nicht derselbe wie zwischen zwei und drei sein muss. Sie sind also trotz des Vorliegens von Zahlen nicht intervallskaliert.

Folglich sind Aussagen zu einem berechneten Mittelwert von beispielsweise 2,63 weniger eindeutig interpretierbar als es bei quantitativen Merkmalen der Fall wäre. Da es sich um einen Fragebogen handelt, sind die Ausprägungen aller Merkmale außerdem entweder diskret oder im Fall einer stetigen Größe diskretisiert. Dies ermöglicht eine einfachere Handhabung des Datensatzes.

Die untersuchten Merkmale des vorliegenden Datensatzes haben unterschiedliche Skalenniveaus und Merkmalsausprägungen. Jeder befragten Person wird jeweils das Merkmal „PersonID“ zugeordnet. Dieses ändert sich auch bei künftigen Befragungen nicht, sodass eine Vergleichbarkeit auch auf individueller Ebene, also für den einzelnen Merkmalsträger, ermöglicht wird. Für diese Arbeit ist die Beantwortung der Frage, ob eine bestimmte Person im Vorjahr ein Krankenhaus aufgesucht hat, sehr wichtig. Demzufolge ist das entsprechende Merkmal ein zentrales. Es stellt im späteren Verlauf der Analyse überdies die abhängige Zielgröße dar.

Ein weiteres fixes Merkmal stellt das Geburtsjahr einer Person dar. Aus diesem wird das Alter abgeleitet, welches einen großen Einfluss auf die Wahrscheinlichkeit einer Erkrankung oder allgemein eines Krankenhausaufenthaltes hat.[45] Wie stark dieser Einfluss bei der Gruppe der Befragten im Detail ist, wird im Verlauf der Datenanalyse ermittelt werden.

Da das Geschlecht, ein nominalskaliertes qualitatives Merkmal, ein zentrales Tarifmerkmal in der privaten Krankenversicherung ist, stellt es für die Untersuchung der Hospitalisierungswahrscheinlichkeit eine wichtige Rolle dar. Die Ergebnisse der Sekundärliteratur implizieren außerdem, dass Menschen mit einem höherem BMI häufiger krank sind als jene mit einem niedrigerem BMI. Um zu überprüfen, ob dies ebenfalls auf die befragten Merkmalsträger des SOEP-Datensatzes zutrifft, werden also auch die Merkmale Größe und Gewicht berücksichtigt.

[45] Vgl. Saß, A.-C. / Wurm, S. / Ziese, T. (2005), S. 32 f.

Es ist allgemein bekannt und durch zahlreiche Studien medizinisch erwiesen, dass Rauchen bereits bei geringem Konsum von Zigaretten der Gesundheit schadet.[46] Um zu Überprüfen, ob die befragten Raucher auch entsprechend häufiger ins Krankenhaus kommen, wird auch dieses Merkmal analysiert. Auch gilt Alkohol, wenigstens in zu großen Mengen, als ungesund. Entsprechend wird der Konsum von Bier, Wein und Spirituosen betrachtet.

Es ist anzunehmen, dass der aktuelle Gesundheitszustand und die Selbsteinschätzung des momentanen Gesundheitsstaus sowie die Einschränkung der Person im Alltag einen großen Einfluss auf die Wahrscheinlichkeit einer Hospitalisierung haben. In einigen Erhebungsjahren wurden darüber hinaus auch chronische Krankheiten wie Diabetes, Herz- oder Gelenkerkrankungen erfragt. Sie stellen ein erhöhtes Krankheitsrisiko dar, insbesondere im Alter.[47] Aufgrund dessen werden diesbezügliche Merkmale in die statistische Analyse einbezogen.

Neben den genannten Merkmalen, von denen erwartet wird, dass sie den Gesundheitszustand bei starker Ausprägung negativ beeinflussen, werden außerdem die Fragen nach Sport und gesundheitsbewusster Ernährung behandelt. Diese liegen jeweils als ordinalskalierte qualitative Merkmale vor und wurden von der Häufigkeit von Sport bzw. der Intensität von gesunder Ernährung zu Ausprägungen der Skala „1 bis 4“ umgewandelt.

Im Zuge der Darstellung der gesundheitsbeeinflussenden Faktoren auf Basis von Sekundärliteratur ist ersichtlich geworden, dass auch der sozioökonomische Status eines Menschen eine zentrale Rolle spielt. Dieser wird häufig durch das Bildungsniveau, den Beruf und das Einkommen bestimmt. Da der Beruf als Merkmal grundsätzlich nominalskaliert ist und keine brauchbaren Daten zum Bildungsniveau vorliegen, wird der Sozialstatus dahingehend berücksichtigt, dass das Monatsgehalt einer Person betrachtet wird.

Aufgrund der Tatsache, dass zwischen vielen der untersuchten Merkmale gewisse Wechselwirkungen herrschen, werden neben den einzelnen Merkmalen auch Korrelationen in die statistische Datenanalyse einbezogen. So ist beispielsweise naheliegend, dass die Merkmale Gesundheitszustand und Gesundheitsbedenken positiv miteinander korreliert sind. Dies muss im Rahmen der Auswertung der einzelnen Merkmale berücksichtigt werden.

---

[46] Vgl. Hackshaw, A. et al. (2018), S. 1
[47] Vgl. Saß, A.-C- / Wurm, S. / Ziese, T. (2005), S. 31

3.2. Darstellung ausgewählter Merkmale

Viele historische Verläufe einzelner Merkmale sowie bestimmte Quoten wurden bereits im Rahmen zahlreicher Studien ermittelt. Ziel dieses Kapitels ist es deshalb, einerseits zu überprüfen, wie repräsentativ die Gesamtheit der Merkmalsträger des SOEP ist. Andererseits werden weitere Merkmale beschrieben und Entwicklungen anschaulich dargestellt. Es werden hierbei jedoch noch keine, über die augenscheinlich erkennbaren, Hospitalisierungswahrscheinlichkeiten ermittelt.

Alle Datenauswertungen des Sozio-oekonomischen Panels werden mithilfe der Programmiersprache R durchgeführt. R bietet eine breite Palette statistischer und grafischer Techniken, welche sehr individuell nutz- und anpassbar sind. Dadurch hat der Nutzer ein hohes Maß an Freiheit bezüglich der Gestaltung seiner Arbeit. Besonders größere Datenmengen, zu welchen der SOEP-Datensatz ebenfalls gehört, sind mithilfe von R benutzerfreundlich und effektiv analysierbar.[48]

Überdies kann R um zahlreiche Pakete erweitert werden. Diese befinden sich im sogenannten „Comprehensive R Archive Network“ (CRAN) und müssen einmalig installiert werden. Dabei wird der Download und die Installation in der Regel mittels Befehl angestoßen, sodass in R die weiteren Schritte eigenständig durchführt werden, indem auf die Internet-Adresse der R Foundation zugegriffen wird.[49]

Um einen besseren Überblick über den R-Code und die zahlreichen Funktionen sowie Plots zu erhalten, wird die Open-Source-Edition des Programms RStudio verwendet. Neben einer Konsole für Befehlseingaben gibt es hier einen Editor zum Schreiben von R-Skripten. Darüber hinaus wird ein Workspace angeboten, in welchem die Umgebung und der Verlauf der Eingaben angezeigt wird. Zur Anschaulichkeit enthält RStudio außerdem einen Data- und Plotviewer sowie einen Package Manager zur Übersicht der installierten Pakete.

Jährlich werden aus rund 15.000 Haushalten über 25.000 Menschen ab dem Alter 17 befragt.[50] Die Auswertung des gesamten Datensatzes ergibt, dass die Anzahl der im Jahr 2016, dem aktuellsten Jahr, befragten Personen bei 29.178 lag. Insgesamt lässt sich ein Anstieg der interviewten Personen erkennen. Die Entwicklung der Anzahl der Befragten ist in der folgenden Abbildung dargestellt:

---

[48] Vgl. The R Foundation
[49] Vgl. Sawitzki, G. (2008), S. I-48
[50] Vgl. Deutsches Institut für Wirtschaftsforschung (c)

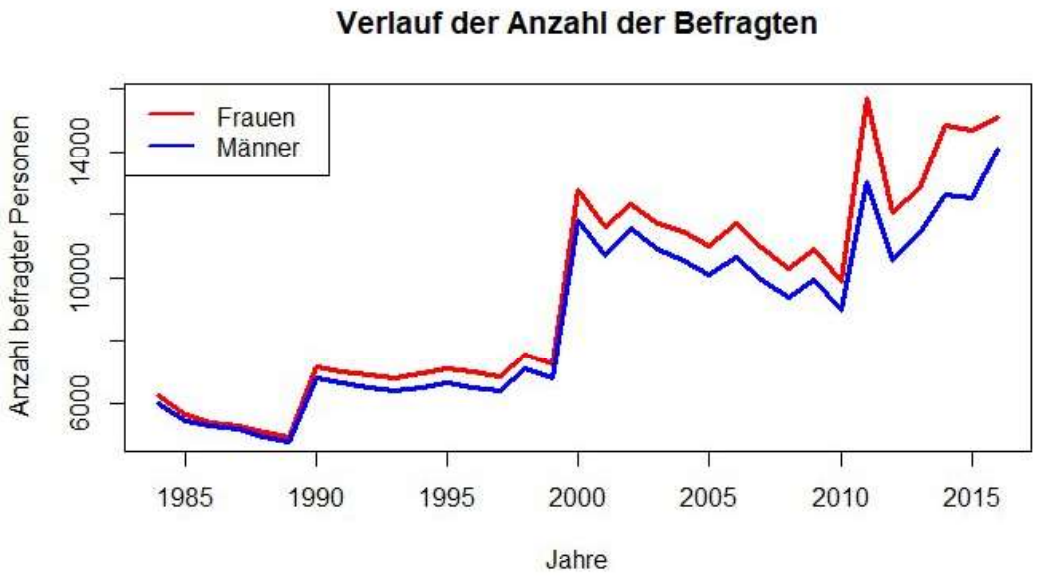

Abbildung 2: Anzahl befragter Personen von 1984 bis 2016

Seit der ersten Befragung im Jahr 1984 ist die Anzahl der Merkmalsträger deutlich gestiegen. Dem Datensatz ist zu entnehmen, dass im Jahr 1984 die Zahl der Befragten 12.290 betrug, 1989 waren es sogar lediglich 9.710 Menschen. Die Anzahl der Frauen ist stets höher als die der Männer. So beträgt der Anteil der Frauen im Jahr 2016 beispielsweise rund 51,67%. Die Aufteilung zwischen Männern und Frauen ist grundsätzlich jedoch als relativ ausgewogen anzusehen, insbesondere wenn man bedenkt, dass der Anteil der Frauen in Deutschland grundsätzlich etwas überwiegt.[51]

Darüber hinaus ist eine hohe Volatilität erkennbar. Dies hat verschiedene Gründe. Im Jahr 1984, also rund fünf Jahre vor dem Mauerfall, wurden lediglich Bürger Westdeutschlands befragt. Beim SOEP werden Mitglieder verschiedener Stichproben immer wieder befragt. Aufgrund der Tatsache, dass die Stichproben mit der Zeit schrumpfen und sich z.B. politische Rahmenbedingungen ändern, kommen auf der anderen Seite neue Stichproben hinzu.

Größere Änderungen der Gesamtanzahl der Merkmalsträger gab es in den Jahren 1990, 2000 und um das Jahr 2011 herum. Im Jahr 1990 kam nach der deutschen Wiedervereinigung eine Stichprobe aus Ostdeutschen hinzu, zur Jahrtausendwende wurde das Panel um eine sogenannte Auffrischungs-Stichprobe erweitert. Ab dem Jahr 2009 kamen mehrere Stichproben hinzu, im Jahr 2013 überdies eine Migrati-

[51] Vgl. Statistisches Bundesamt (2017c), S. 26

ons-Stichprobe.[52] Eine detaillierte Betrachtung der Entwicklung der Stichproben zeigt die Abbildung „Stichprobenentwicklung des Sozio-oekonomischen Panels“ (s. Anhang, Abb. 18).

Dass der Datensatz des Sozio-oekonomischen Panels die Gesamtheit der in Deutschland lebenden Menschen gut widerspiegelt, zeigt auch die Verteilung der Befragten nach Alter und Geschlecht. Die folgende Abbildung zeigt die absoluten Häufigkeiten der einzelnen Altersklassen innerhalb des Datensatzes:

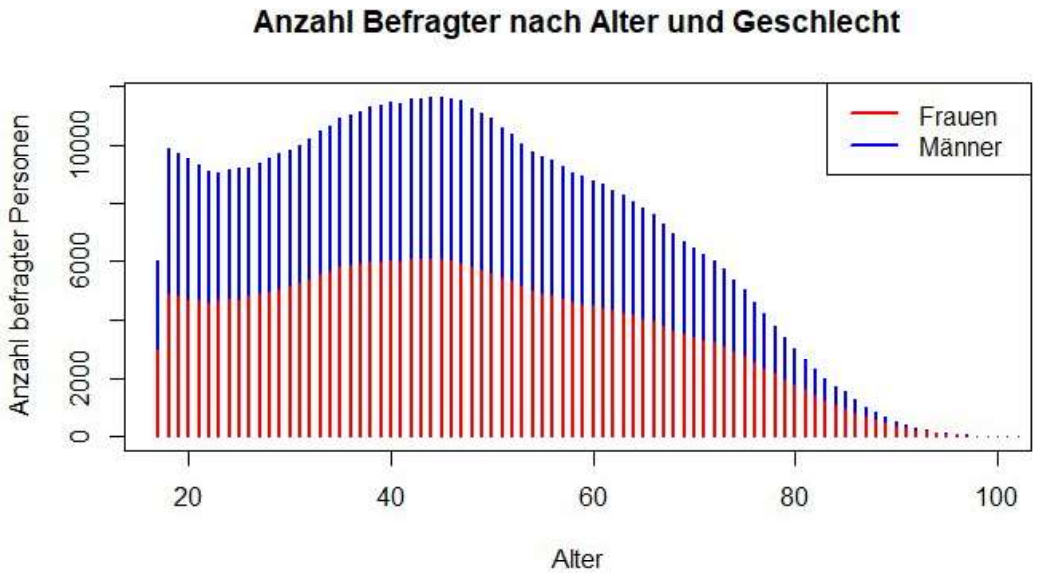

Abbildung 3: Anzahl Befragter nach Alter und Geschlecht, 1984 bis 2016

Man erkennt die für Deutschland typische „Urnenform“, die durch die spezielle Altersstruktur Deutschlands zustande kommt. Es ist hierbei jedoch zu beachten, dass die Abbildung nicht ein bestimmtes Jahr zeigt, sondern den gesamten Zeitraum der Datenerhebung, also 1984 bis 2016. Dadurch wird eine Volatilität aufgrund zu geringer Datenmengen verringert. Ein Vergleich der einzelnen Jahre 1984 und 2016 findet sich im Anhang wieder (s. Anhang, Abb. 19 und Abb. 20).

Betrachtet man das Durchschnittsalter der Befragten, erkennt man den Trend, dass die Bevölkerung Deutschlands im Laufe der Zeit älter wird. War das Durchschnittsalter des Kollektivs im ersten Jahr der Befragung noch rund 42,12 Jahre, so lag es 32 Jahre später bei 46,24 Jahren. Auffällig ist, dass das Maximum im Jahr 2012 liegt und, nachdem der Mittelwert kontinuierlich angestiegen ist, 51,75 Jahre beträgt. Ab 2013 sank das Durchschnittsalter wieder. Es liegt die Vermutung nahe, dass dies

[52] Vgl. Deutsches Institut für Wirtschaftsforschung (d)

unter anderem auf die Erweiterung des Kollektivs um Migrations-Stichproben zurückzuführen ist. Insgesamt liegt Altersdurchschnitt der Befragten oberhalb des Bundesdurchschnitts, was nicht zuletzt daran liegt, dass nur Personen ab 17 Jahren befragt wurden.

Ein zentrales Merkmal, der Gesundheitszustand, ist ein deutliches Indiz dafür, dass eine Person in naher Zukunft ein Krankenhaus aufsuchen muss. Sport gilt, sofern man es nicht übertreibt, im Allgemeinen als gesund. Deshalb werden im folgenden der Gesundheitszustand und die Häufigkeit von Sport gegenübergestellt, um zu überprüfen, ob Sportler tatsächlich einen besseren Gesundheitszustand aufweisen. Dabei werden jedoch noch keine Berechnungen durchgeführt, sondern lediglich folge Abbildung untersucht:

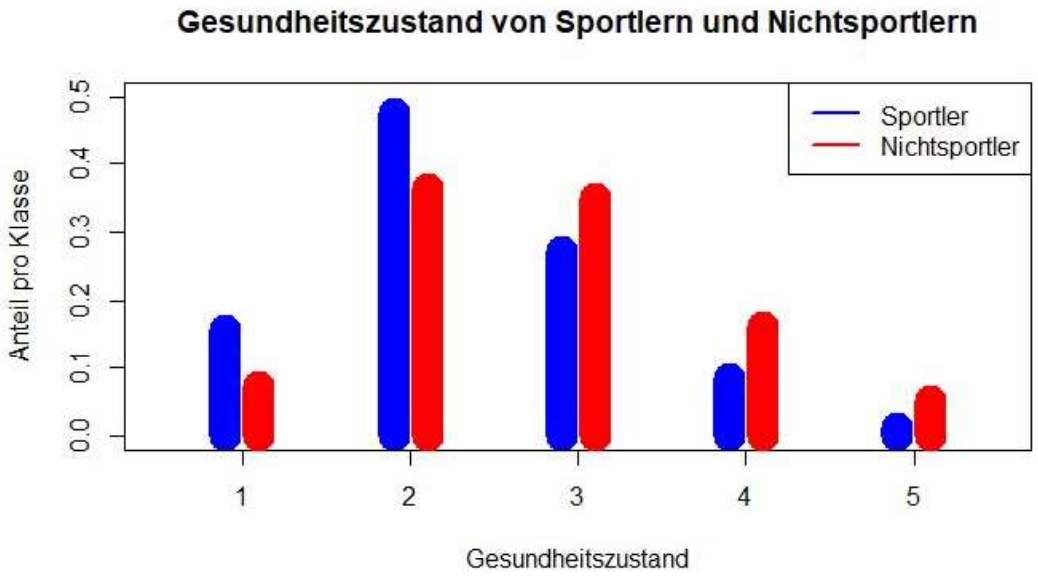

Abbildung 4: Gesundheitszustand von Sportlern und Nichtsportlern

Der Gesundheitszustand ist in fünf Kategorien gegliedert. Die erste Kategorie hat die Merkmalsausprägung „sehr gut“, die letzte „schlecht“. Bezüglich der sportlichen Aktivitäten wurden die Befragten gefragt, wie häufig sie einer solchen nachgehen. Zu der Kategorie „Sportler“ werden in den obigen Abbildung jene Personen gezählt, die jede Woche oder jeden Monat Sport treiben. „Nichtsportler“ sind hingegen diejenigen, die angegeben haben, selten oder nie Sport zu machen.

Zwar gibt es Sportler, die einen schlechten und Nichtsportler, welche einen sehr guten Gesundheitszustand aufweisen. Dennoch lässt sich anhand der Abbildung eine deutliche Tendenz erkennen, dass Menschen, die regelmäßig einer sportlichen

Aktivität nachgehen, gesünder sind als Nichtsportler. Diese Beobachtung deckt sich ebenfalls mit den Erkenntnissen der Medizin.[53]

Bei dem Gesundheitszustand handelt es sich um ein ordinalskaliertes qualitatives Merkmal. Indem man den einzelnen Zuständen ganze Zahlen zuschreibt, lassen sich jedoch jeweils für Sportler und Nichtsportler Mittelwerte berechnen. So ergibt sich für die Gruppe der Sportler ein Wert von 2,32 und für die Nichtsportler ein Schnitt von 2,75. Der Einfluss von Sport auf die Gesundheit eines Menschen kann somit offenbar nicht das einzige Kriterium sein, obgleich ein Zusammenhang angenommen werden kann.

Wie in vorangegangenen Kapiteln erläutert, stellt bei der Frage nach der Gesundheit auch das Raucherverhalten der Menschen ein wichtiges Merkmal dar. Aufgrund dessen wurde der Datensatz ebenfalls diesbezüglich ausgewertet. Obwohl das Krankheitsrisiko durch Rauchen erheblich erhöht und der Gesundheitszustand verschlechtert wird, lässt sich diese Tatsache nicht durch die Daten aus dem Pool der befragten Personen bestätigen. Die Frage nach dem Raucherverhalten wurde erst ab dem Jahr 2002 Teil des Fragebogens, sodass die Anzahl der seitdem abgegebenen gültigen Antworten 261.249 beträgt.

Es ist nicht festzustellen, ob die Antworten der Raucher bezüglich des Gesundheitszustands fehlerhaft beantwortet worden sind, oder ob im Falle dieser Stichproben tatsächlich ein Zusammenhang zwischen dem Rauchverhalten und dem Gesundheitszustand nicht besteht. Dies jedenfalls ist die Beobachtung, die sich aus dem Datensatz des SOEP ergibt. Denn der Anteil der Raucher ist in nahezu jeder Kategorie des Gesundheitszustandes genauso hoch wie der der Nichtraucher (s. Anhang, Abb. 21)

Betrachtet man die Entwicklung der Quote der Raucher von 2002 bis 2016, so erkennt man, dass sie sehr stark schwankt. Zur Untersuchung der Quote wurden die Personen je nach Alter jeweils den Teilkollektiven „unter 50 Jahre“ bzw. „über 50 Jahre“ zugeordnet. Beide Teilkollektive sind etwa gleich groß, die Summe der Merkmalsträger beträgt folglich in beiden Fällen rund 130.000 Befragte. Die folgende Abbildung zeigt die beiden entsprechenden Verläufe der jährlichen Anteile der Raucher an der Gesamtheit der Befragten:

[53] Vgl. Bös, K. / Woll, A. (2017), S. 1

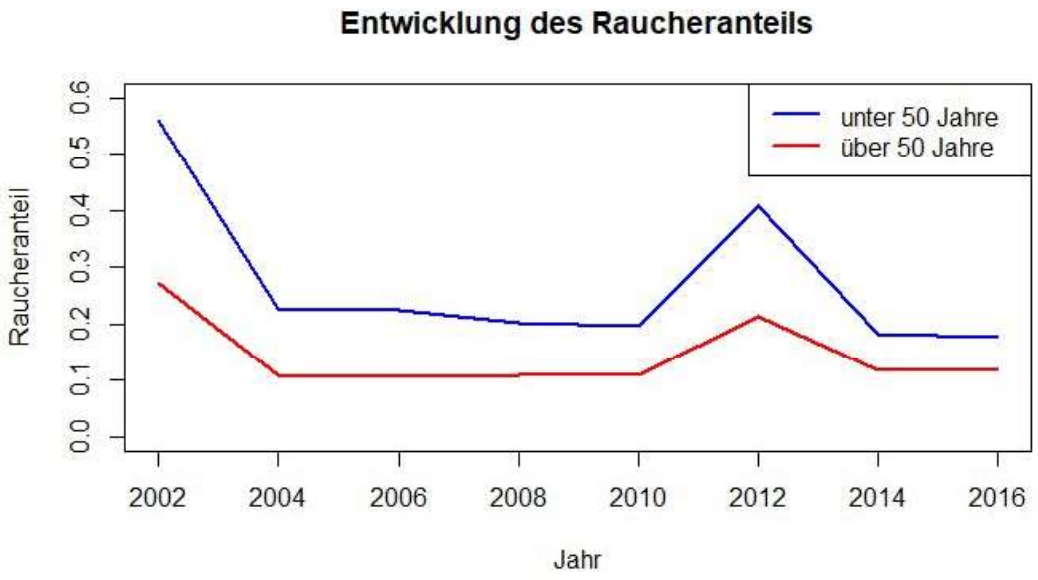

Abbildung 5: Entwicklung des Raucheranteils nach Alter, 2002 bis 2016

Zunächst fällt auf, dass gemäß dem Datensatz ältere Menschen weniger rauchen als jüngere. Dies deckt sich mit den Ergebnissen des Mikrozensus des Jahres 2017 und liegt daran, dass vor allem ab dem Alter von 65 Jahren die Anzahl rauchender Personen stark sinkt.[54] Das Absinken lässt sich wiederum teilweise durch die erhöhte Mortalität von Rauchern, insbesondere Langzeit-Rauchern, erklären. Demnach ist die Anzahl der Raucher im Rentenalter geringer, da die Lebenserwartung der Raucher geringer ist.

Darüber hinaus stellt man beim Betrachten der Abbildung in einigen Jahren starke Veränderungen der Raucheranteile fest. In den Jahren 2002 und 2012 war die Quote, insbesondere bei der jüngeren Gruppe, extrem hoch. Sie betrug 2002 beinahe 56%, was einen ungewöhnlich hohen Anteil darstellt. Für die unter 50-jährigen liegt die Anzahl der Raucher in den meisten Jahren zwischen rund 3.000 und 5.000 Personen, während etwa 12.000 bis 18.000 der Antworten auf Nichtraucher entfallen. In den Jahren 2002 und 2012 hingegen, gaben zwar rund 5.000 bzw. 3.000 Menschen an, Raucher zu sein, die Anzahl der Nichtraucher in diesen Jahren betrug jedoch jeweils nur etwa 4.000 Personen.

Durch Einflüsse der Stichprobenentwicklung lassen sich solche erheblichen Unterschiede nicht erklären, vor allem nicht in einem solch plötzlichen und erheblichen Ausmaß. Nicht zuletzt aufgrund der weniger starken Veränderung der Raucher im Vergleich zur extrem starken Veränderung bei den Nichtrauchern liegt wiederum die

[54] Vgl. Statistisches Bundesamt (2017d)

Vermutung nahe, dass die Daten, zumindest im Bezug auf die Nichtraucheranteile, lückenhaft sind.

Aufgrund der Ergebnisse aus der Sekundärliteratur sind auch die Beobachtungen der jeweiligen Body-Mass-Indizes der befragten Personen von Relevanz. Jemand mit einem BMI ab 25 gilt laut Weltgesundheitsorganisation als übergewichtig und hat in der privaten Krankenversicherung einen nicht unerheblichen Risikozuschlag zu zahlen. Deshalb wird einerseits die Verteilung der BMI über die einzelnen Altersklassen, andererseits der Anteil derjenigen Personen mit einem BMI von über 25 betrachtet. Dies ist in der folgenden Abbildung für die Jahre 2002 und 2016 dargestellt:[55]

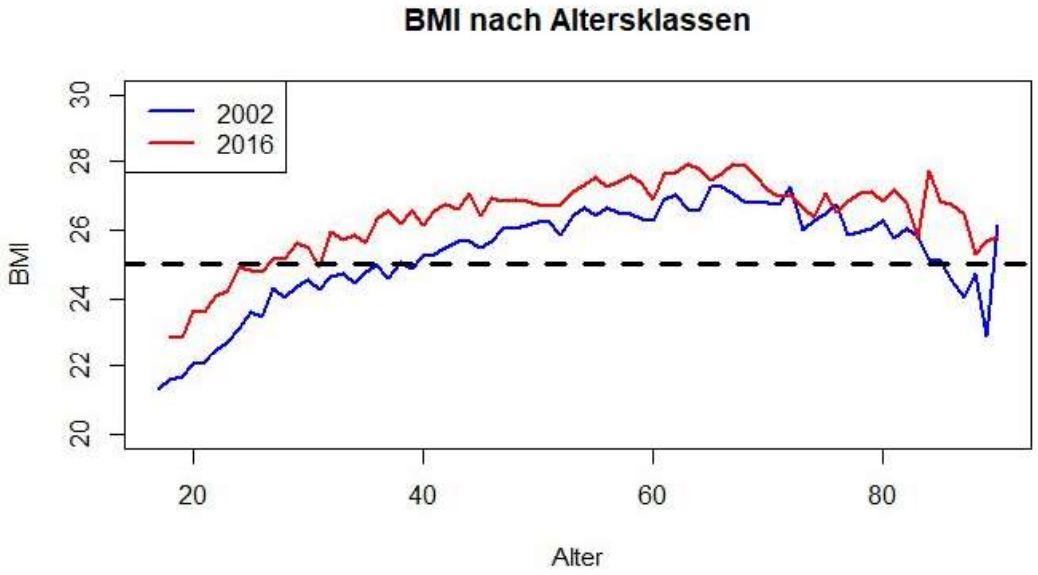

Abbildung 6: BMI nach Altersklassen, 2002 und 2016

Grundsätzlich nimmt der Mittelwert des BMI bei steigendem Alter zu. Es fällt jedoch auf, dass ab einem gewissen Alter das Gewicht einer Person im Mittel offenbar wieder sinkt. Dieses beträgt im Fall des vorliegenden Datensatzes etwa 60 Jahre. Es ist überdies anzumerken, dass die Abbildung lediglich Befragte zwischen 17 und 90 Jahren berücksichtigt, da die Anzahl derjenigen Merkmalsträger, die älter als 90 Jahre alt sind, sehr gering ist. Ein Einbezug dieser Personen hätte folglich einen stark volatilen Verlauf der Kurve mit geringer Aussagekraft zur Folge.

Bei Betrachtung der Abbildung fällt darüber hinaus auf, dass die Mittelwerte des BMI im Jahr 2002 noch deutlich niedriger waren als im Jahr 2016. Die Durchschnittswer-

[55] *Konkrete Werte der Altersklassen 20, 40, 60 und 80 finden sich im Anhang, Tab. 9 wieder.*

te liegen für 2016 bei nahezu jeder Altersgruppe über den Werten von rund 14 Jahren zuvor. Dies betrifft auch den Anteil derjenigen Personen, die einen BMI von über 25 und damit ein Übergewicht aufweisen.

Lag der Anteil der übergewichtigen Befragten des Datensatzes im Jahr 2002 noch bei rund 47,95%, waren im Jahr 2016 bereits etwa 55,14% übergewichtig. Der Anteil adipöser Menschen, also jener Personen mit einem BMI von 30 oder höher, lag in diesen Jahren bei 13,11% bzw. 19,80%. Männer wiesen hierbei im Mittel einen höheren BMI als Frauen auf. Dass ein Anstieg der Quote Übergewichtiger erkennbar ist, zeigen auch andere Studien, die sich mit der Gewichtsentwicklung Deutschlands befassen.[56]

Zum Abschluss dieses Kapitels wird im folgenden die Gehaltsstruktur des analysierten Datensatzes betrachtet. Laut Sekundärliteratur ist sie ein wichtiger Indikator für den sozioökonomischen Status, welcher wiederum einen Einfluss auf den Gesundheitszustand einer Person haben kann. Aufgrund dessen wird das Gehalt in der folgenden Abbildung in Bezug zum aktuellen Gesundheitszustand gesetzt:

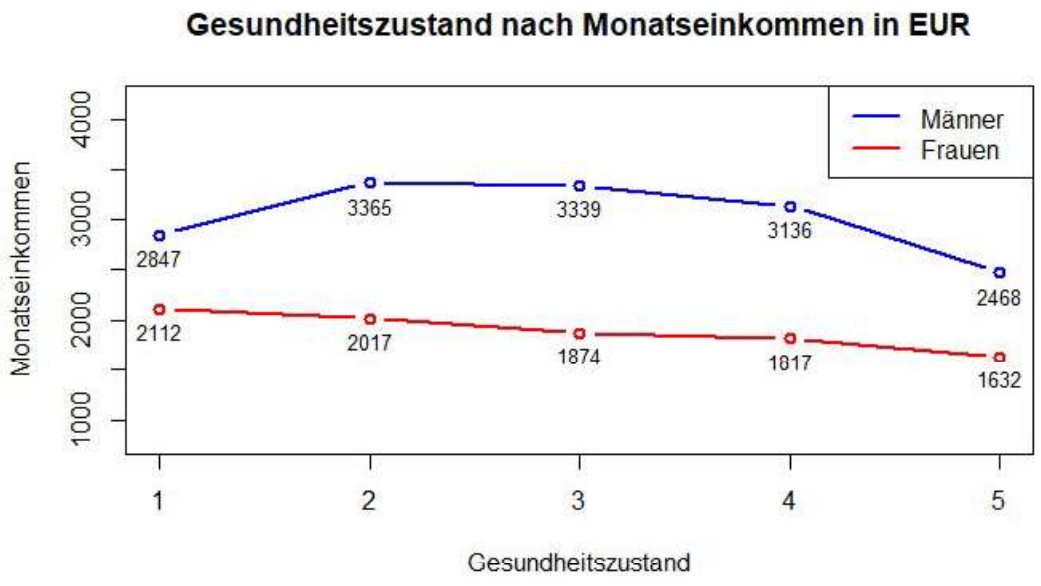

Abbildung 7: Gesundheitszustand in 2016 nach Monatseinkommen in EUR

Zunächst fällt auf, dass es auch auf den SOEP-Datensatz zutrifft, dass es mitunter große Unterschiede des Gehalts zwischen Männern und Frauen gibt. So verdienten Männer in 2016 im Schnitt rund 3.249 Euro, wohingegen Frauen durchschnittlich nur 1.954 Euro Gehalt erhielten. Zwar ist das durchschnittliche Gehalt bei den Männern

[56] Vgl. Mensink, G.B.M. et al. (2013), S. 787 ff.

auch dann höher, wenn Männer und Frauen die gleiche Tätigkeit ausüben, es existiert also ein sog. bereinigter Gender Pay Gap.[57] Ein weiterer Grund für die Differenzen liegt jedoch auch darin, dass Frauen häufiger die Rolle der Hausfrau übernehmen als Männer, entsprechend kein Gehalt verdienen. Dies wiederum senkt beim vorliegenden Datensatz insgesamt den Durchschnitt des Monatseinkommen auf nur etwa 60% des Mittelwertes der Männer.

Auf Basis der Abbildung kann überdies eine weitere Feststellung gemacht werden. Der Gesundheitszustand scheint offenbar mit dem Monatseinkommen einer Person zusammenzuhängen. Sowohl bei der Gruppe der Männer als auch bei den Frauen ist erkennbar, dass Menschen mit höherem Einkommen einen besseren Gesundheitszustand aufzuweisen scheinen als jene mit niedrigerem Einkommen. Abgesehen von einer Ausnahme, nimmt die Gesundheit mit sinkendem Monatsgehalt kontinuierlich ab.

Lediglich diejenigen Männer, die in einem sehr guten Gesundheitszustand sind, haben offenbar ein geringeres Monatseinkommen als die meisten anderen befragten Männer. Eine Erklärung hierfür kann anhand der vorliegenden Daten nicht gefunden werden. Auch die Verteilung der einzelnen Werte weist keine Besonderheit auf, die Standardabweichung der ersten Gruppe unterscheidet sich nicht grundlegend von den Standardabweichungen der Gruppen mit anderen Gesundheitszuständen.

[57] Statistisches Bundesamt (2017e), S. 19

3.3. Statistische Analyse anhand ausgewählter Merkmale

Nachdem zahlreiche, für die Hospitalisierungswahrscheinlichkeit relevante, Merkmale betrachtet worden sind, wird im Folgenden mithilfe statistischer Analysen die Wahrscheinlichkeit einer Hospitalisierung ermittelt. Hierfür werden mit dem Programm RStudio die entsprechenden Merkmale selektiert und gefiltert, um sie daraufhin in verschiedene Regressionsanalysen einfließen zu lassen. Auf Basis dieser Methoden sollen die ermittelten Wahrscheinlichkeiten eine Prädiktion ermöglichen, ob eine Person im Folgejahr ins Krankenhaus kommen wird.

Die angewendeten Regressionen sind einerseits die lineare Regression (LinReg), andererseits die etwas komplexere logistische Regression (LogReg), welche auch Logit-Modell genannt wird.[58] Dabei werden zunächst die mathematischen Grundlagen der Analysen erläutert, woraufhin dann die Berechnungen in RStudio folgen. Diese werden sowohl grafisch dargestellt als auch auf ihre Aussagekraft hin überprüft. Hierbei erfolgt eine Plausibilisierung einerseits durch Beurteilung nach dem „gesunden Menschenverstand“, andererseits durch als Standard anerkannte mathematische Tests und Berechnungen.

3.3.1. Vorgehen und Rahmenbedingungen

Bei der statistischen Analyse des SOEP-Datensatzes werden im Wesentlichen zu zwei Teilkollektiven Regressionsmodelle erstellt. Es handelt sich dabei zum einen um das Basismodell, zum anderen um das Alternativmodell. Die hierfür relevanten Merkmale sind unterschiedlich, sodass ein Vergleich zwischen den beiden Modellen gezogen werden kann, um zu ermitteln, welche Merkmale einen größeren Einfluss auf die Krankenhauswahrscheinlichkeit haben. Überdies existiert ein drittes, didaktischen Zwecken dienendes, lineares Modell, bei welchem lediglich das Alter einer Person auf die Hospitalisierungswahrscheinlichkeit bezogen wird.

Die benötigten Merkmale für das Basismodell und das Alternativmodell stammen jeweils aus verschiedenen Jahren. So liegen dem Training des Basismodells Daten aus dem Jahr 2006 zugrunde, wohingegen das Alternativmodell auf Daten aus 2009 zugreift. Das Testing dieser Modelle bezieht sich jeweils zunächst auf das Folgejahr, also 2007 bzw. 2010. Außerdem wird das Basismodell auch im Jahr 2010 getestet, um eine bessere Vergleichbarkeit mit dem Alternativmodell zu ermöglichen.

[58] Vgl. Wagner, S. (2015)

Aufgrund der Tatsache, dass für die Überprüfung, ob tatsächlich ein Krankenhausaufenthalt stattgefunden hat, das Folgejahr des untersuchten Jahres betrachtet wird, fließen insgesamt alle Jahre von 2006 bis 2011 in die Gesamtanalyse ein. Diese Überprüfung ist an dieser Stelle möglich, da es sich um vergangenheitsbezogene Daten handelt. Damit die Modelle auch für Aussagen über möglicherweise bevorstehende Krankenhausaufenthalte genutzt werden können, wird beim Anwenden des Basismodells auf das Jahr 2010 nicht nur ein Vergleich zum Alternativmodell vorgenommen, sondern überdies auch überprüft, ob es zeitkonsistent ist. Hierzu erfolgt ein Vergleich mit den Werten aus dem Training dieses Modells.

In diesem Kapitel wird zunächst die einfache lineare Regression dargestellt und anhand des Alters überprüft, inwiefern sich dieses Modell eignet, die Wahrscheinlichkeit zu bestimmen, dass eine Person im Folgejahr ins Krankenhaus kommen wird. Im Rahmen dessen werden zunächst allgemein die Grundlagen dieser Regression erläutert. Anschließend werden die mithilfe von R durchgeführten Berechnungen dargestellt, bei welcher tatsächliche Daten der Stichprobe verwenden werden.

Die einfache lineare Regression findet dann Anwendung, wenn ein gerichteter Zusammenhang zwischen zwei Variablen unterstellt wird. Dabei handelt es sich einerseits um die abhängige Variable X, andererseits um die unabhängige Variable Y. Die Variable X ist demnach ein vorliegender Wert, der wiederum einen Einfluss darauf hat, welchen Wert Y annehmen wird. Entsprechend wird X auch Einflussvariable, Y auch Zielvariable genannt.[59] Ein klassisches Beispiel hierfür stellt der Zusammenhang von Preis und Absatzmenge dar.

Um ein lineares Regressionsmodell aufstellen zu können, wird eine bestimmte Mindestanzahl bereits vorliegender Beobachtungswerte benötigt. Je mehr beobachtete Werte aus der Stichprobe vorhanden sind, desto genauer bzw. passender kann die daraus abgeleitete Regressionsgerade aufgestellt werden. Bei der Aufstellung der linearen Regressionsgeraden wird unter anderem das Ziel verfolgt, die Ergebnisse einer Stichprobe zu verallgemeinern und in Form eines linearen Zusammenhangs darzustellen.

Hierzu betrachtet man zunächst die einzelnen vorliegenden Beobachtungswerte der Stichprobe. Für eine grobe Übersicht geschieht dies häufig in Form einer sogenannten Punktewolke bzw. eines Streudiagramms. Durch die Verteilung der einzelnen

[59] Vgl. Frost, I. (2018), S. 5

Beobachtungswerte wird dem Beobachter eine Einschätzung der Art des Zusammenhangs ermöglicht. Man sollte dann zum Beispiel erkennen können, in welche Richtung eine mögliche Trendlinie verlaufen könnte, also etwa, ob steigende X-Werte auch steigende Y-Werte nach sich ziehen, oder nicht. Überdies sind Ausreißer in einer Punktwolke besonders gut zu beobachten. Bei diesen Punkten passen die beobachteten X- und Y-Werte nicht zu den anderen Paaren.

Da bei dem linearen Regressionsmodell ein linearer Zusammenhang zwischen abhängigen und unabhängigen Variablen vermutet wird, stellt sich die Geradengleichung und damit die Ermittlung der abhängigen Variable Y_i entsprechend wie folgt dar[60]:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i$$

Der Wert Y_i bezieht sich auf die i-te Person des Datensatzes. Bei dieser Geradengleichung stellt β_0 den Ordinatenabschnitt und β_1 den Steigungsfaktor der Geraden dar. Überdies enthält die Gleichung die Störgröße ε_i, welche auch als Fehlervariable bezeichnet wird. Sie ist für jeden einzelnen Punkt individuell. Durch sie wird ausgedrückt, dass die Punkte nicht alle genau auf der Geraden liegen, sondern um sie streuen.[61] Für die Störgröße ε_i gibt es folgende Annahmen:

$$(I)\ E(\varepsilon_i) = 0$$

$$(II)\ Var(\varepsilon_i) = \sigma^2$$

$$(III)\ Cov(\varepsilon_i, \varepsilon_j) = 0 \ \text{für}\ i \neq j$$

Der Erwartungswert einer einzelnen Störgröße beträgt somit 0 und die Standardabweichung σ. Außerdem sind die einzelnen Störgrößen voneinander unabhängig, sie sind also auch unkorreliert.[62] Wird an die Störgrößen zusätzlich die Forderung gestellt, dass sie normalverteilt mit $E(\varepsilon_i) = 0$ und $Var(\varepsilon_i) = \sigma^2$ sind, so handelt es sich um ein Regressionsmodell mit normalverteilten Störungen. Dann gilt insbesondere:[63]

$$E(y_i) = E(\beta_0 + \beta_1 * x_i + \varepsilon_i) = \beta_0 + \beta_1 * x_i$$

Aufgrund der mitunter hohen Anzahl der Beobachtungswerte stellt sich die Frage, welche Geradengleichung die Beziehung zwischen den Variablen X und Y insge-

[60] Vgl. Schlittgen, R. (2013), S. 7
[61] Vgl. Frost, I. (2018), S. 3
[62] Vgl. Frost, I. (2018), S. 6
[63] Vgl. Schlittgen, R. (2013), S. 7 f.

samt am besten darstellt. Hierzu müssen die jeweiligen Parameter der Gleichung geschätzt werden. Bei der Methode der kleinsten Quadrate wird das Ziel verfolgt, die Koeffizienten so zu wählen, dass die Residuen $\hat{\varepsilon}_i = y_i - \hat{y}_i$ möglichst klein sind. Die Zielfunktion der linearen Regression lautet demnach:

$$\sum_{i=1}^{I} \hat{\varepsilon}_i^{\,2} \rightarrow \min_{\hat{\beta}_0, \hat{\beta}_1}$$

Es wird hierbei jeweils das Quadrat der Residuen verwendet. Dies hat einerseits zur Folge, dass verhindert wird, dass negative Residuen sich mit positiven kompensieren. Anderseits werden durch das Quadrieren große Abweichungen stärker gewichtet. Die Koeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ werden als Kleinste-Quadrate-Schätzfunktionen bzw. -Schätzwerte bezeichnet und es ergeben sich durch Extremwertbestimmung folgende Formeln:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

Demnach ist also die geschätzte Steigung der Regressionsgeraden die Kovarianz von X und Y geteilt durch die Varianz des Merkmals X. Der Ordinatenabschnitt ist dann die Differenz zwischen dem Mittelwert der beobachteten Y-Werte und dem Produkt aus $\hat{\beta}_1$ und dem Mittelwert der beobachteten X-Werte. Folglich lassen sich die zu schätzenden Y-Werte mithilfe fester, unabhängiger X-Werte durch die allgemeine Regressionsfunktion bestimmen:[64]

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$$

Nachdem die Grundlagen der einfachen linearen Regression dargelegt worden sind, wird sie im Folgenden auf den vorliegenden SOEP-Datensatz angewendet. Wie für die anderen Regressionen auch, ist hier die abhängige Variable Y das Merkmal des Krankenhausaufenthalts. Es handelt sich hierbei um ein dichotomes Merkmal, da die Merkmalsausprägung nur entweder „ja“ oder „nein“ lauten kann.

Die Ausprägung dieses Merkmals bezieht sich überdies auf das Folgejahr, damit diesbezüglich tatsächlich eine Vorhersage getroffen werden kann. Bei den Berechnungen zur einfachen linearen Regression werden im Folgenden die ermittelten Y-

[64] Vgl. Schlittgen, R. (2013), S. 9

Werte als Wahrscheinlichkeiten für einen Krankenhausaufenthalt einer Person angesehen.

Als Prädiktor, also als unabhängige Variable, die auf die Wahrscheinlichkeit eines Krankenhausaufenthaltes einwirkt, dient in diesem Fall das Merkmal „Alter“. Damit eine gewisse Vergleichbarkeit mit dem Basismodell vorgenommen werden kann, ist der dem Modell zugrunde liegende Datensatz der gleiche wie der des Basismodells. Hierbei wurden neben dem Alter auch weitere Merkmale selektiert und das Erhebungsjahr auf 2006 eingegrenzt. Diese Merkmale werden im Folgekapitel näher thematisiert.

Infolgedessen erhält man ein Teilkollektiv, welches 17.755 befragte Personen enthält. Die Personen dieses Kollektivs sind zwischen 17 und 97 Jahren alt und es liegt überdies zu allen jeweils eine Information vor, ob sie im Jahr 2007 ein Krankenhaus aufgesucht haben. Dies trifft auf 2.101 Personen, also rund 11,83% der Befragten, zu. Hieraus wiederum ergibt sich mithilfe von Berechnungen in R folgende lineare Regressionsgerade:

$$\hat{y}_i = -0{,}0117761 + 0{,}0026629 * x_i$$

Die geschätzte Wahrscheinlichkeit, dass eine Person im Folgejahr, also im Jahr 2007, ins Krankenhaus kommen wird, wird durch die Variable $\hat{Y}_i$ dargestellt. Aus der Tatsache, dass ein linearer Zusammenhang mit einer positiven Steigung besteht, folgt, dass ältere Menschen im Folgejahr mit höherer Wahrscheinlichkeit ein Krankenhaus aufsuchen werden als jüngere.

Setzt man also beispielsweise das Alter 17 in die obige Geradengleichung ein, so ergibt sich ein Wert von rund 3,35%. Für die 97-jährige Person hingegen, welche die älteste des Teilkollektivs ist, resultiert durch Einsetzen des X-Wertes eine geschätzte Wahrscheinlichkeit von rund 24,65%. Der Verlauf der Regressionsgeraden stellt sich wie folgt dar:

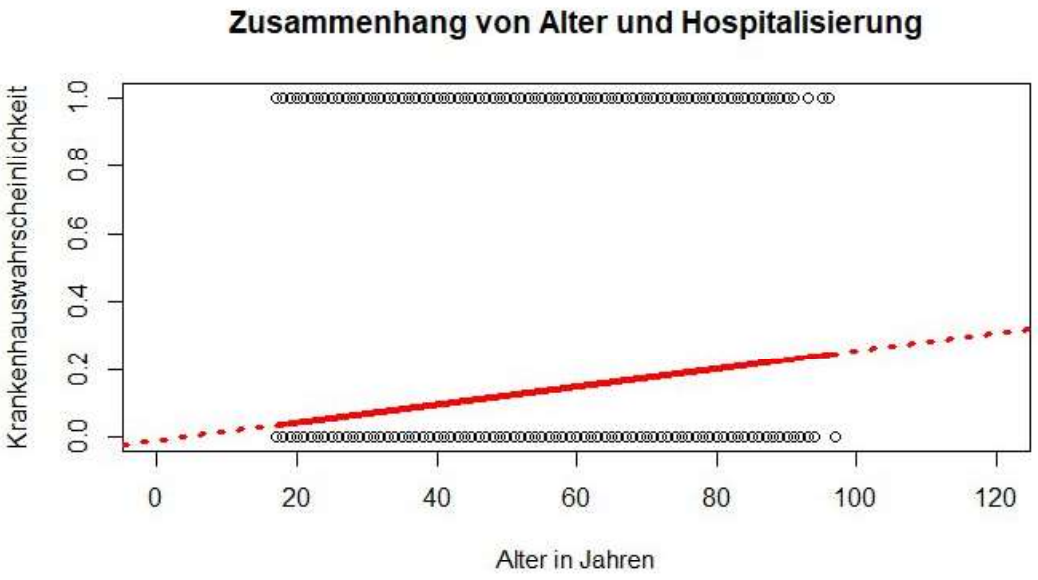

Abbildung 8: Zusammenhang von Alter und Hospitalisierung, LinReg

Die zahlreichen Punkte bzw. Ausprägungen in dieser Abbildung stellen jeweils die Beobachtungen für die einzelnen Personen dar, ob sie in 2007 ein Krankenhaus aufgesucht haben, oder nicht. Da es sich bei der Frage nach einem Krankenhausaufenthalt um ein kategoriales bzw. nominales Merkmal handelt,[65] wurden die Merkmalsausprägungen „ja“ und „nein“ in die reellen Zahlen 1 und 0 transformiert. Folglich ist in diesem Fall auch keine typische Punktewolke zu erkennen, deren Punkte um die Gerade verteilt sind.[66] Stattdessen wird die Regressionsgerade durch die beiden Ausprägungen 1 und 0 beeinflusst.

Die durchgezogene Linie stellt die ermittelte Krankenhauswahrscheinlichkeit der Personen des Datensatzes dar. Die gepunktete Linie hingegen zeigt die theoretischen geschätzten $\hat{Y}_i$ bei Einsetzen anderer X_i an. Hierdurch erkennt man außerdem deutlich den Ordinatenabschnitt der Regressionsgerade, welcher sich knapp unterhalb von 0 befindet.

Beim näheren Betrachten des einfachen linearen Regressionsmodells wird schnell klar, dass das Alter allein als Prädiktor zu wenig Aussagekraft hat. So war beispielsweise der 97-jährige, welcher laut Modell die höchste Wahrscheinlichkeit einer Hospitalisierung aufwies, im Jahr 2007 nicht im Krankenhaus, während mehr als 15% aller 17-jährigen in 2007 stationär behandelt wurden. Kinder haben nach diesem Modell eine extrem geringe Hospitalisierungswahrscheinlichkeit. So wird für

[65] Vgl. Hartung, J. / Elpelt, B. (2007), S. 19
[66] Für e*in Beispiel mit einer klassischen Punktewolke siehe Anhang, Abb. 22.*

Kinder, die das vierte Lebensjahr noch nicht vollendet haben, sogar eine negative Wahrscheinlichkeit prognostiziert, sofern man die berechneten Y-Werte als Wahrscheinlichkeiten interpretiert.

Es stellt sich also die Frage, wie gut die Regressionsgerade tatsächlich zum Abschätzen der abhängigen Variablen ist. Die sogenannte Residuenstreuung kann unterschiedlich groß sein. So ist es theoretisch möglich, dass Stichproben mit geringerer Streuung zu der gleichen Regressionsgerade führen wie jene mit großer Streuung. Die daraus resultierende Qualität der Schätzungen ist folglich jeweils unterschiedlich zu bewerten.

Es existieren zahlreiche Ansätze, die Güte eines Modells zu bewerten. Darüber hinaus werden ständig neue Gütemaße entwickelt. Um jedoch eine Vergleichbarkeit der verschiedenen Modelle gewährleisten zu können, werden im Rahmen dieser Arbeit zwei gängige Gütemaße betrachtet. Dabei handelt es sich einerseits um das sog. Bestimmtheitsmaß R^2, andererseits wird für jedes Modell jeweils die sog. „Area under the curve“ (AUC) berechnet. Beide Gütemaße werden im Verlauf der Analyse erläutert und zu einer standardisierten Überprüfung der Modellgüte herangezogen.

Das Bestimmtheitsmaß gibt an, wie gut die abhängigen Variablen geeignet sind, die Varianz der abhängigen Variablen zu erklären. Der Wertebereich des R^2 lautet $0 \leq R^2 \leq 1$. Ein perfektes Modell hätte ein R^2 von 1 bzw. 100%, wohingegen das Bestimmtheitsmaß eines völlig unbrauchbaren Modells 0 bzw. 0% betragen würde. Die Berechnung des R^2 erfolgt, indem die durch die unabhängigen Variablen erklärte Streuung mit der gesamte Streuung ins Verhältnis gesetzt wird. Die gesamte Streuung besteht neben der genannten erklärten außerdem aus der unerklärten Streuung:[67]

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Der erste Summand stellt hierbei die erklärte Streuung dar, der zweite die unerklärte. Daraus folgt, dass die Berechnung des Bestimmtheitsmaßes wie folgt vorzunehmen ist:

$$R^2 = \frac{\sum_{i=1}^{I}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2}$$

[67] Vgl. Schlittgen, R. (2013), S. 6

Im Fall der einfachen linearen Regression, welche die Merkmale Alter und Krankenhausaufenthalt im Folgejahr zugrunde legt, ergibt sich das folgende Bestimmtheitsmaß:

$$R^2 = \frac{36{,}20}{1852{,}38} \approx 0{,}019542 \mathrel{\widehat{=}} 1{,}9542\%$$

Dieser Wert bedeutet, das nur rund 1,9542% der Varianz der abhängigen Variablen Y, also die Frage, ob jemand im Folgejahr im Krankenhaus war, durch die unabhängige Variable X, also durch das Alter einer Person, erklärt werden kann. Dieser Wert ist sehr gering, da die Abweichungen der tatsächlichen Y-Werte von 1 bzw. 0 oft stark von den Prädiktionen bzw. deren Mittelwert von rund 11,83% abweichen. Man sagt auch, es handelt sich hierbei um einen sog. „poor model fit".[68]

Da hier die einfache lineare Regression zum Vorhersagen einer binären abhängigen Variable offensichtlich nicht geeignet scheint, wird im Folgenden die logistische Regression vorgestellt. Sie findet vor allem dann Anwendung, wenn das Ziel der statistischen Analyse die Vorhersage einer Wahrscheinlichkeit für das Eintreten eines bestimmten Falls ist. Es gibt für die Beurteilung der Hospitalisierungswahrscheinlichkeit mehrere Gründe, die für eine Anwendung eines Logit-Modells sprechen.

Zum einen wird beim linearen Regressionsmodell ein linearer Zusammenhang zwischen Prädiktor und abhängiger Variable unterstellt. Dies hat zur Folge, dass die Wahrscheinlichkeiten sowohl Werte unter 0 als auch über 1 annehmen können, wie im obigen Fall eines Kleinkindes beispielhaft gezeigt worden ist. Dies bedeutet, dass der Wertebereich auf [0,1] beschränkt werden muss, wodurch wiederum einige Prädiktionen wegfallen können. Die logistische Regression ist bei kategorialen abhängigen Variablen mit dichotomer Ausprägung besser als die lineare Regression geeignet.[69]

Darüber hinaus sind die Schätzwerte, die aus der linearen Regression resultieren, stellenweise nicht plausibel. Im Beispiel der Abhängigkeit der Hospitalisierungswahrscheinlichkeit vom Alter erscheint etwa ein linearer Zusammenhang nicht sinnvoll. Die Hospitalisierungswahrscheinlichkeit für einen Menschen steigt mit zunehmendem Alter nicht linear an. So haben Menschen hohen Alters beispielsweise ein höheres Risiko, ins Krankenhaus zu kommen, als ein linearer Zusammenhang vermuten ließe. Etwa wird einer Person des Alters 80 Jahre eine Wahrscheinlichkeit

[68] Vgl. Pflieger, V. (2014)
[69] Vgl. Schlittgen, R. (2013), S.215

von rund 20,13% zugeschrieben, wohingegen der Anteil der 80-jährigen, die in 2007 stationär behandelt wurden, bei rund 26,79% lag.

Das Vorgehen bei einer Regressionsanalyse wird oft in fünf Einzelschritten beschrieben.[70] Der erste Schritt ist die Modellformulierung. Hier werden die relevanten Größen selektiert, also die Prädiktoren und im Fall der logistischen Regression die Response-Variable Y. Bei manchen Merkmalen kann es sinnvoll sein, sie zusammenzufassen. Notwendig ist dies vor allem dann, wenn die abhängige Variable nicht nur zwei Ausprägungen hat, sondern ein multinomialer Fall vorliegt, zusätzlich also beispielsweise die Ausprägung „vielleicht" existiert.

Darüber hinaus können Hypothesen aufgestellt werden. Dabei ist jedoch zu beachten, dass nicht zwischen der abhängigen und der unabhängigen Variablen eine direkte Beziehung besteht, sondern vielmehr zwischen der unabhängigen Variablen und der Eintrittswahrscheinlichkeit für die binäre abhängige Variable. Im zweiten Schritt erfolgt die Schätzung der logistischen Regressionsfunktion. Hierzu wird im Folgenden eine Herleitung der allgemeinen logistischen Funktion vorgenommen.

Die logistische Regression gehört zu den strukturprüfenden Verfahren. Da die untersuchten Y-Werte binär bzw. dichotom sind, handelt es sich hier um eine binäre logistische Regression. Entsprechend gilt allgemein für die Wahrscheinlichkeiten der beiden Merkmalsausprägungen die folgende Beziehung:

$$P(Y = 1) = 1 - P(Y = 0)$$

Die Wahrscheinlichkeit, dass jemand im Folgejahr ein Krankenhaus aufsuchen muss, beträgt demnach:[71]

$$\pi(P(Y = 1|x)$$

Die Wahrscheinlichkeit für einen Krankenhausaufenthalt ist also eine bedingte Wahrscheinlichkeit, dessen Bedingung die jeweilige Ausprägung der unabhängigen Variable ist. Dies gilt grundsätzlich auch bei der linearen Regression. Um nun einen nichtlinearen Zusammenhang beobachten zu können, werden bei der logistischen Regression die latenten Variablen z_i als Linearkombination der unabhängigen Variable angenommen. Diese werden auch Logits genannt und sind zunächst nicht empirisch beobachtbar, dienen jedoch als Bindeglied zwischen unabhängiger und abhängiger Variable. Die Gleichung für z_i lautet wie folgt:

[70] Vgl. Backhaus, K. Et al. (2016), S. 289 ff.
[71] Vgl. Backhaus, K. et al. (2016), S. 284

$$z_i = \beta_0 + \beta_1 * x_i + \varepsilon_i$$

Es fällt auf, dass diese Formel zunächst eine sehr starke Ähnlichkeit mit dem linearen Modell aufweist. Auch z_i kann folglich Werte außerhalb des Bereichs von [0,1] annehmen. Um zu verhindern, dass dies auch für die Prädiktionen gilt, wird die logistische Funktion verwendet. Hierzu zunächst wird der Quotient aus der Wahrscheinlichkeit für einen Eintritt und der Wahrscheinlichkeit für einen Nichteintritt gebildet. Dieser Wert wird auch Odds genannt:[72]

$$Odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{\pi}{1-\pi}$$

Im Beispielfall beträgt dieser Wert folglich 0,1342. Theoretisch können die Odds auch einen Wert von über 1 annehmen, falls die Anzahl der Eintritte höher ist als die der Nichteintritte. Damit der Wertebereich [0,1] ist, werden die Odds im nächsten Schritt logarithmiert.[73] Dieser logarithmierte Wert wiederum entspricht den Logits. Demnach lässt sich die logistische Funktion durch die folgende Gleichung herleiten:

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = z = \beta_0 + \beta_1 * x$$

Infolge des Logarithmierens werden β_0 und β_1 zu sog. Logit-Koeffizienten. Durch Auflösen dieser Gleichung nach $P(Y=1)$ erhält man schließlich die logistische Funktion:[74]

$$P(Y=1) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Bei der linearen Regression geschieht das Schätzen der Modellparameter durch die Kleinste-Quadrate-Methode. Aufgrund der Tatsache, dass die logistische Regression jedoch nicht linear ist, wird die Parameterschätzung hier mithilfe der Maximum-Likelihood-Methode vorgenommen. Hierbei werden die Koeffizienten der logistischen Funktion so bestimmt, dass die realisierten Daten maximale Plausibilität erlangen.[75] Nachfolgend werden die Grundlagen der Maximum-Likelihood-Methode erläutert, spätere Parameterschätzungen werden jedoch mithilfe von R durchgeführt.

---

[72] Vgl. Wentura, D. / Pospeschill, M. (2015), S. 61

[73] *Eine Übersicht des Zusammenhangs zwischen Odds und Logits findet sich im Anhang, Abb. 24 wieder.*

[74] *Eine ausführliche Umformung der Gleichung befindet sich im Anhang (S. 71).*

[75] Vgl. Backhaus, K, et al. (2016), S. 305

Bei der Maximum-Likelihood-Methode wird das Ziel verfolgt, dass für eine Person die Wahrscheinlichkeit möglichst groß sein soll, falls das Ereignis eintritt, also Y=1 ist. Für den Fall, dass Y=0 ist, soll die Wahrscheinlichkeit entsprechend gering ausfallen. Demnach soll im Fall der logistischen Regression das folgende Produkt einen möglichst hohen Wert annehmen:

$$P_i(y) = \left(\frac{1}{1+e^{-z_i}}\right)^{y_i} * \left(1 - \frac{1}{1+e^{-z_i}}\right)^{1-y_i}$$

Aufgrund der Annahme der Unabhängigkeit der Personen, kann die gemeinsame Wahrscheinlichkeit für alle Personen als Produkt der Einzelwahrscheinlichkeiten ausgedrückt werden. Dies ergibt die Likelihood-Funktion, welche ebenfalls zu maximieren ist:[76]

$$L = \prod_{i=1}^{I} \left(\frac{1}{1+e^{-z_i}}\right)^{y_i} * \left(1 - \frac{1}{1+e^{-z_i}}\right)^{1-y_i} \rightarrow max!$$

Für die praktische Berechnung ist es sinnvoll, die Wahrscheinlichkeiten jeweils zu logarithmieren, sodass man statt des Produktes eine Summe berechnen kann. Dies ist deshalb möglich, da es auch nach dem Logarithmieren dieselben Parameter sind, die maximiert werden sollen.[77] Überdies ist der Logarithmus eine streng monoton steigende Funktion. Die Exponenten y_i und $(1 - y_i)$ werden vor den Logarithmus gezogen. Folglich ergibt sich die sog. Log-Likelihood-Funktion:

$$LL = \sum_{i=1}^{I} \left(y_j * \ln\left(\frac{1}{1+e^{-z_i}}\right)\right) + \left((1-y_i) * \ln\left(\frac{1}{1+e^{-z_i}}\right)\right)$$

Im Fall der Verwendung des Alters als einzigen Prädiktoren ergibt sich durch Anwendung der Maximum-Likelihood-Methode für die Logits bzw. z_i die folgende Gleichung:

$$z_i = -3{,}350414 + 0{,}02597049 * x_i$$

Wie bei der einfachen linearen Regression auch, liegt hier also zunächst ein linearer Zusammenhang vor. Es ist jedoch zu beachten, dass diese Gleichung nicht die geschätzte Wahrscheinlichkeit für einen Krankenhausaufenthalt darstellt. Stattdessen ist dieser Wert nun in die zuvor hergeleitete logistische Funktion einzusetzen. Daraus ergeben sich für jede Person mit ihren individuellen Merkmalsausprägungen jeweils Wahrscheinlichkeiten für einen Krankenhausaufenthalt im Folgejahr.

---

[76] Vgl. Backhaus, K. et al. (2016), S. 306
[77] Vgl. Dreo, R. (2014), S. 29

Nach dem Aufstellen der logistischen Regressionsfunktion werden in einem dritten Schritt die einzelnen Regressionskoeffizienten interpretiert.[78] Aufgrund der Tatsache, dass zunächst nur das Alter als Prädiktor angenommen wird, fließt auch nur diese Größe in die Ermittlung der Hospitalisierungswahrscheinlichkeit ein. Die Folge ist, dass für Menschen gleichen Alters entsprechend dieselbe Wahrscheinlichkeit geschätzt wird. Ein Verlauf der logistischen Funktion in Abhängigkeit der Logits zeigt die folgende Abbildung:

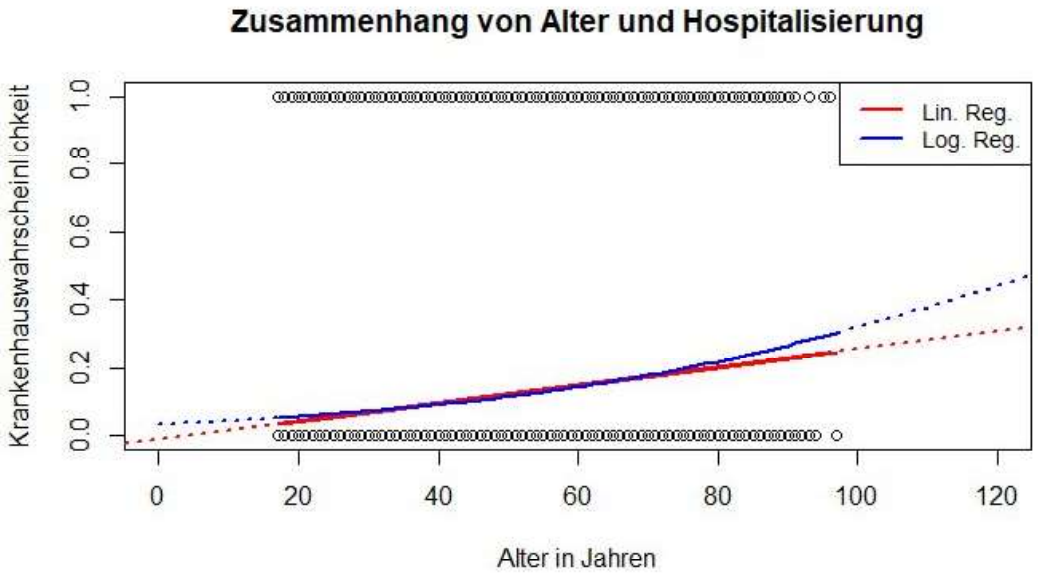

Abbildung 9: Zusammenhang von Alter und Hospitalisierung, LinReg u. LogReg

Wie man dieser Abbildung entnehmen kann, hat die logistische Funktion einen anderen Verlauf als die Kurve der linearen Regression. Auch hier stellen die durchgezogenen Linien diejenigen Prädiktionen dar, die für die Personen des Kollektivs gelten. Die gepunkteten Linien zeigen die hypothetischen Wahrscheinlichkeiten eines Krankenhausaufenthaltes für jedes andere beliebige Alter.

Anders als bei der linearen Regression, nimmt die Kurve der logistischen Funktion keine Werte unter 0 oder über 1 an, sondern nähert sich diesen beiden Werten je nach Merkmalsausprägung lediglich immer näher an. Dieser Zusammenhang wird dann umso deutlicher, je größer die Spanne der Regressoren, in diesem Fall also der Altersklassen, ist (s. Anhang, Abb. 23).

Nachdem man jeweils die Wahrscheinlichkeiten für die einzelnen Personen ermittelt hat, kann eine Klassifizierung der Werte erfolgen. Da es das Ziel der logistischen

[78] Vgl. Backhaus, K. et al. (2016), S. 308

Regression ist, Aussagen darüber zu treffen, ob das abhängige Ereignis eintreten wird, ist es sinnvoll, eine Schwelle zu setzen, ab welcher ein Eintritt angenommen wird. Dadurch erhält beispielsweise der durchführende Versicherer die Möglichkeit, auf diejenigen Versicherungsnehmer zuzugehen, die diesen Schwellenwert überschreiten, um daraufhin gezielte Präventivmaßnahmen einzuleiten.

Wie hoch dieser Schwellenwert ist, ist zum einen davon abhängig, um was für eine abhängige Variable es sich handelt. Zum anderen ist es vor allem eine unternehmenspolitische Entscheidung, die auch durch die Risikoaffinität des Entscheiders beeinflusst wird. Durch „klassisches“ Runden würde eine Beurteilung des Eintritts der abhängigen Variable zunächst wie folgt aussehen:

$$Y_i = \begin{cases} 1 & falls\ P_i(Y=1) \geq 0{,}5 \\ 0 & falls\ P_i(Y=1) < 0{,}5 \end{cases}$$

Aufgrund der Tatsache, dass es sich bei den Logits um die logarithmierten Odds handelt, kann die Beurteilung des Schwellenwerts auch bereits mit diesen erfolgen. Entsprechend stellt sich der Zusammenhang dann wie folgt dar:

$$Y_i = \begin{cases} 1 & falls\ z_i \geq 0 \\ 0 & falls\ z_i < 0 \end{cases}$$

Da bei der logistischen Regressionsanalyse die einzelnen Koeffizienten der Logits noch in die Regressionsfunktion eingesetzt werden müssen, um tatsächlich Prädiktionen zu erhalten, kann man Änderungen dieser Koeffizienten nicht direkt interpretieren. Es sind jedoch Richtungen des Einflusses des Regressors erkennbar. Ein Vermindern bzw. Erhöhen der Regressionskonstante β_0 führt zu einer horizontalen Verschiebung der Funktion nach rechts bzw. links.[79] Es besteht also eine gewisse Ähnlichkeit zur linearen Regression, bei welcher eine Änderung der Konstante ebenfalls zu einer Lageverschiebung führt.

Diese Ähnlichkeit gilt auch für die Regressionskoeffizienten der Logits. Sie besteht dahingehend, dass sie auch bei der logistischen Funktion zum Steigungsverhalten der Kurve beitragen. Ist $\beta_1 = 0$, so ist die Kurve nicht mehr s-förmig, sondern eine horizontale Linie, sodass die Wahrscheinlichkeiten für alle Beobachtungen 50% betragen. Die Steigung der Regressionskurve ist umso stärker, je größer das β_1 ist. Liegt ein negativer Regressionskoeffizient vor, wird die Kurve umgedreht, es handelt sich dann folglich um einen abfallenden Verlauf.

[79] Vgl. Backhaus, K. et al. (2016), S. 308

Der vierte Schritt bei der Durchführung einer logistischen Regression ist die Prüfung des Gesamtmodells. Hierzu werden einzelne Werte auf Plausibilität sowie die Güte des Modells überprüft. Im Fall des Alters als einzigen Prädiktor ergäbe sich bei Anwenden der zuvor erläuterten 50%-Schwelle, dass lediglich für diejenigen Personen ein Krankenhausaufenthalt im Folgejahr vorhergesagt wird, welche das 129. Lebensjahr bereits vollendet haben. Ein Schwellenwert von 50% ergäb folglich keinen Sinn.

Aus diesem Grund existieren verschiedene Möglichkeiten, einen passenden Schwellenwert festzulegen. Auf diese Möglichkeiten wird im Folgekapitel eingegangen. Betrachtet man nun für die logistische Regression die Wahrscheinlichkeit für einen 80-jährigen, im Jahr 2007 in ein Krankenhaus zu kommen, erhält man folgendes Resultat:

$$P(Y=1) = \frac{1}{1+e^{-(-3{,}350414+0202597049*80)}} = \frac{1}{1+e^{3{,}57075}} \approx 0{,}218783 \mathrel{\widehat{=}} 21{,}8783\%$$

Verglichen mit dem Ergebnis der linearen Regression von rund 20,13%, stellt die mithilfe der logistischen Regression ermittelte Wahrscheinlichkeit für einen Krankenhausaufenthalt von rund 21,88% eine bessere Schätzung dar. Der Vergleich mit dem Anteil derjenigen 80-jährigen, die tatsächlich in 2007 stationär behandelt wurden, zeigt jedoch, dass auch die einfache logistische Regression für eine gute Schätzung nicht auszureichen scheint. Aufgrund dessen wird im Folgenden auch für das logistische Regressionsmodell das Bestimmtheitsmaß ermittelt.

Durch die Eigenschaft der abhängigen Variable, nicht metrisch, sondern kategorial zu sein, ist die Berechnung des klassischen Bestimmtheitsmaßes für die logistische Regression nicht möglich. Aus diesem Grund wird oftmals ein sog. Pseudo-Bestimmtheitsmaß verwendet. Auch bei dem verwendeten Pseudo-Bestimmtheitsmaß liegt der Wertebereich zwischen 0 und 1. Außerdem bedeuten auch hier hohe Werte eine gute, niedrige Werte eine schlechte Anpassung.

Die sich daraus ergebenden Werte sind jedoch nicht genauso zu interpretieren wie die des herkömmlichen Bestimmtheitsmaßes, sie erklären also nicht den Anteil der erklärten Streuung an der gesamten Streuung. Stattdessen wird hier das Verhältnis zweier Wahrscheinlichkeiten betrachtet. Dabei handelt es sich um die Likelihood einerseits eines sog. 0-Modells, andererseits eines vollständigen Modells.[80]

[80] Vgl. Backhaus, K. et al. (2016), S. 317

Es existieren zahlreiche Ansätze, von denen vor allem drei in der Praxis angewendet werden. Dabei handelt es sich um das Cox & Snell R^2, das Nagelkerkes R^2 und das McFadden R^2. Als Bestimmtheitsmaß für logistische Regression dient in dieser Arbeit das McFadden R^2. Bei diesem wird ein Quotient der Log-Likelihoods gebildet. Dabei wird das vollständige Modell durch ein Nullmodell geteilt:

$$R^2{}_{McF} = 1 - \left(\frac{LL_v}{LL_0}\right)$$

Weicht der Wert des 0-Modells kaum von dem des vollständigen ab, dann ist der Quotient nahezu 1, das McFadden R^2 folglich nahezu 0. Dies bedeutet, dass je größer der Unterschied der Werte der beiden Modelle ist, desto mehr kann das Modell erklären. Entsprechend ist dann das Pseudo-Bestimmtheitsmaß größer. Für den Fall des Zusammenhangs von Alter und Krankenhauswahrscheinlichkeit ergibt sich mithilfe von R ein McFadden-Bestimmtheitsmaß von:

$$R^2{}_{McF} = 1 - \left(\frac{-6281}{-6456}\right) = 0{,}027050 \mathrel{\widehat{=}} 2{,}7050\%$$

Zwar kann das Bestimmtheitsmaß der linearen Regression nicht direkt mit dem McFadden R^2 verglichen werden. Dennoch stellen beide einen Wertebereich von 0 bis 1 dar, sodass Tendenzen erkennbar sind. Mit einem Wert von rund 1,95% ist das R^2 der linearen Regression kleiner als das Pseudo-Bestimmtheitsmaß der logistischen Regression von rund 2,70%. Doch auch, wenn beim McFadden R^2 bereits kleinere Werte eine gute Modellanpassung bedeuten können,[81] ist auch das Pseudo-Bestimmtheitsmaß der logistischen Regression gering. Dies lässt den Schluss zu, dass ein Modell, welches nur das Alter als unabhängige Variable vorsieht, unzureichend ist.

Nachdem das Gesamtmodell überprüft worden ist, folgt in einem letzten Schritt in der Regel die Prüfung der Merkmalsvariablen. Hierbei wird überprüft, ob ein bestimmter Prädiktor einen signifikanten Einfluss auf die abhängige Variable hat. Bei der linearen Regression verwendet man in der Regel den sog. t-Test. Bei der logistischen Regression hingegen, werden vornehmlich zwei andere Tests angewendet. Einerseits handelt es sich um den sog. Wald-Test, andererseits um den Likelihood-Ratio-Test. [82]

[81] Vgl. Backhaus, K. et al. (2016), S. 317
[82] Vgl. Backhaus, K. et al. (2016), S. 320

Bei der Überprüfung der Signifikanz der einzelnen Koeffizienten werden in dieser Arbeit jedoch ausschließlich die ausgegebenen Werte von R betrachtet, weshalb auf genauere Definitionen des Wald- und Likelihood-Ratio-Tests an dieser Stelle verzichtet wird. Dabei handelt es sich zum einen um den z-Wert, welcher den Quotienten aus geschätztem Koeffizienten β_k und empirischer Standardabweichung des Merkmals darstellt. Zum anderen wird der noch interessantere p-Wert betrachtet.

Der p-Wert, auch „prob-value“, ist eine Prüfgröße zur Bestimmung der Signifikanz. Er stellt das Maß für die Plausibilität der Nullhypothese H_0 dar. Es gilt, dass der p-Wert desto kleiner ist, je signifikanter das Merkmal innerhalb des Modells ist.[83] Da sowohl beim linearen als auch beim logistischen Modell nur das Alter berücksichtigt wird, sind die p-Werte bei beiden Regressionen 0.

3.3.2. Basismodell

Da offenbar das Alter allein nicht ausreicht, um brauchbare Aussagen über die Krankenhauswahrscheinlichkeit einer Person im Folgejahr treffen zu können, wird im Folgenden das Basismodell vorgestellt. Dieses berücksichtigt mehrere unabhängige Variablen, welche einen Einfluss auf die Gesundheit haben können und somit als Prädiktoren geeignet sind. Die einzelnen Merkmale werden jeweils dargestellt. Darüber hinaus wird erläutert, warum bestimmte Merkmale nicht in das Basismodell aufgenommen worden sind.

Überdies werden auch diejenigen Änderungen dargelegt, die sich für die lineare und die logistische Regression aus der Hinzunahme weiterer Merkmale ergeben. Außerdem wird ein weiteres Standard-Gütemaß vorgestellt und die Ergebnisse der Basisdaten und deren Interpretation sowohl für die lineare als auch für die logistische Regressionsanalyse dargestellt.

Aufgrund der Tatsache, dass nun mehr Merkmale als nur das Alter zur Prädiktion verwendet werden, handelt es sich nicht mehr um eine einfache, sondern um eine sog. multiple Regression. Hier ist die abhängige Variable Y also nicht mehr nur noch von X abhängig, sondern von den k Merkmalen $X_1, X_2, \ldots, X_k$. Aus diesem Grund wird auch die Funktionsgleichung der einfachen linearen Regression erweitert. Jeder

[83] Vgl. Dreo, R. (2014), S. 33

Regressor hat einen individuellen Einfluss auf den Wert von Y. Folglich lautet die allgemeine Gleichung der multiplen linearen Regression:

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_k * x_{ik} + \varepsilon_i$$

Auch hier stellt i wieder den Laufindex für die jeweilige Person dar, da Y für jede Person einen eigenen Wert annimmt. Im Fall der einfachen linearen Regression kommt es, insbesondere bei klassierten Größen wie dem Alter, häufig vor, dass der Prädiktor für zwei verschiedene Personen identisch ist. Folglich sind dann auch die Y-Werte identisch. Durch die Berücksichtigung mehrerer Variablen, kann bei der multiplen linearen Regression eine individuellere Schätzung pro Person erfolgen.

Die einzelnen Koeffizienten werden auch als partielle Regressionskoeffizienten bezeichnet. Die Bezeichnung partiell wird deshalb verwendet, weil der Regressionskoeffizient β_k für jedes k die Veränderung des Erwartungswerts von Y misst, die sich durch eine Änderung von X_k um eine Einheit ergibt, wenn die Werte der anderen Prädiktoren gleich bleiben.[84] Auch für die multiple lineare Regression gelten einige Modellannahmen:

$$(I)\ \ y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_k * x_{ik} + \varepsilon_i$$

$$(II)\ \ E(\varepsilon_i) = 0$$

$$(III)\ \ Var(\varepsilon_i) = \sigma^2$$

$$(IV)\ \ Cov(\varepsilon_i, \varepsilon_j) = 0 \quad f\ddot{u}r\ i \neq j$$

Die erste Annahme sieht vor, dass X auf Y einen linearen Einfluss hat. Liegt also etwa ein quadratischer Zusammenhang zwischen X und Y vor, so ist diese Annahme verletzt. Darüber hinaus wird bei der linearen Regression angenommen, dass der Erwartungswert von ε_i den Wert 0 hat, sich der Einfluss der Störgröße im Mittel also aufhebt. Außerdem soll die Varianz der Störgrößen konstant sein, also Homoskedastie vorliegen. Entsprechend gilt die Annahme (III).[85]

Eine weitere Annahme sieht vor, dass die Störgrößen untereinander unkorreliert sind. Dies bedeutet, dass die Kovarianz zwischen zwei Störgrößen 0 ist. Das lineare Regressionsmodell kann in Matrixform mit $y = X\beta + \varepsilon$ geschrieben werden:[86]

---

[84] Vgl. Schlittgen, R. (2013), S. 20
[85] Vgl. Handl, A. / Kuhlenkasper, T. (2017), S. 220
[86] Vgl. Handl, A. / Kuhlenkasper, T. (2017), S. 222

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix},\ x = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$$

Im einfachen linearen Regressionsmodell erfolgt die Parameterschätzung mithilfe der Methoden der kleinsten Quadrate. Die Zielfunktion, welche auch im vorangegangen Kapitel verwendet worden ist, lautet:

$$\sum_{i=1}^{I} (y_i - \beta_0 - \beta_k * x_{ik})^2 \rightarrow min$$

Durch Darstellung dessen als in der Matrixform, kann die Zielfunktion auch wie folgt dargestellt werden:

$$(y - X\beta)' (y - X\beta)$$

Diese Zielfunktion kann nun auch beim Vorliegen mehrerer Prädiktoren verwendet werden, um die Regressionskoeffizienten zu schätzen. Dazu wird eine Hyperebene so durch die Punkte gelegt, dass die Summe der quadrierten Abstände der Punkte von der Hyperebene minimal ist. Durch Extremwertbestimmung ergibt sich somit folglich:[87]

$$\hat{\beta} = (X'X)^{-1}X'y$$

Da die Schätzwerte beim Basismodell mithilfe von R ermittelt werden, wird auf die Parameterschätzung an dieser Stelle nicht weiter eingegangen. Die Interpretation der einzelnen Koeffizienten erfolgt ähnlich wie bei der einfachen linearen Regression. Die Regressionskonstante β_0 wird hier auch als Achsenabschnitt bezeichnet. Aufgrund der Tatsache, dass es nun jedoch mehrere X-Werte gibt, handelt es sich nicht um einen Schnittpunkt einer Geraden mit der Y-Achse. Stattdessen wird die Gleichung in einem mehrdimensionalem Raum beschrieben.[88]

Genau handelt es sich um k+1 Dimensionen. Ist, wie im Fall der einfachen linearen Regression, k=1, so handelt es sich also um ein zweidimensionales System. Auch für k=2 ist der Zusammenhang zwischen dem Y-Wert und den X-Werten noch vorstellbar und mithilfe eines dreidimensionalen Koordinatensystems sogar grafisch darstellbar (s. Anhang, Abb. 25). Bei höheren Dimensionen wird dies jedoch immer komplexer. Sowohl im einfachen als auch im multiplen Fall ist das Bestimmtheits-

[87] Handl, A. / Kuhlenkasper, T. (2017), S. 224
[88] Vgl. Wagner, S. (2014)

maß der linearen Regression als der Anteil der erklärten Variation an der Gesamtvariation definiert.

Auch für die logistische Regression ergeben sich durch das Hinzukommen weiterer Prädiktoren Änderungen. Die Linearkombinationen z_i der unabhängigen Variablen werden, analog zur linearen Regression, erweitert. Demnach ergibt sich bei der multiplen logistischen Regression für die Logits die folgende Gleichung:

$$z_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_k * x_{ik} + \varepsilon_i = \beta_0 + \sum_{k=1}^{K} \beta_k * x_{ik} + \varepsilon_i$$

Die Regressionskonstante sowie die einzelnen Regressionskoeffizienten sind also auch hier nicht personenspezifisch. Da nun mehrere X-Werte vorhanden sind, werden diese neben dem Laufindex i ebenfalls mit dem Laufindex k fortgeführt. Die Aufstellung der logistischen Regressionsfunktion mittels logarithmierten Odds wird daraufhin analog zum einfachen Fall durchgeführt.

Auch die Parameterschätzung findet wie beim einfachen logistischen Regressionsmodell statt, also über die Maximum-Likelihood-Methode. Aufgrund der Tatsache, dass, bedingt durch ihre Charakteristik, für die logistische Regression kein Bestimmtheitsmaß berechnet werden kann, wird auch im multiplen Fall das Pseudo-Bestimmtheitsmaß nach McFadden ermittelt.

Nachdem die notwendigen Anpassungen der beiden Regressionsmodelle erläutert worden sind, werden im Folgenden die einzelnen Schritte der Regressionsanalyse durchgeführt. In einem ersten Schritt, der Modellformulierung, wird zunächst festgelegt, welche Merkmale als Prädiktoren für das Basismodell in Frage kommen. Insgesamt beinhaltet der analysierte Teil des SOEP-Datensatzes 36 Merkmale, von denen eines die abhängige Variable Y darstellt (s. Anhang, Tab. 10).

Die übrigen Merkmale sind auf Basis der Forschungsergebnisse aus der Sekundärliteratur ausgewählte potenzielle Prädiktoren. Es handelt sich dabei teilweise um Grunddaten zu den befragten Personen, teilweise zielen sie auf das Verhalten, insbesondere das Gesundheitsverhalten, der Befragten ab. Darüber hinaus beinhaltet der Datensatz das Erhebungsjahr sowie eine individuelle Personen-ID.

Von diesen 35 Regressoren werden einige sowohl in das Basis- als auch in das Alternativmodell einbezogen. Manche finden sich nur im Basis-, andere ausschließlich im Alternativmodell wieder. Wieder andere werden in keines der Modelle einbezogen. Die Anzahl der im Basismodell beachteten Merkmale beträgt 13. Insgesamt

bleiben also 22 der für die Regressionsanalysen als Prädiktor in Frage kommenden Merkmale im Basismodell unberücksichtigt. Davon sind neun Merkmale Teil des im Folgekapitel behandelten Alternativmodells.

Grund hierfür ist, dass die Merkmale „Gewicht“, „Größe“, „Raucher“ sowie Fragen zum Alkoholkonsum einer Person jeweils alle zwei Jahre erfragt werden. Die für die Analyse relevanten Jahre sind also 2006, 2008 und 2010. Andere, für die Gesundheit und damit die Hospitalisierungswahrscheinlichkeit ebenfalls relevante Merkmale wie „Sport“ oder zahlreiche chronische Krankheiten, können nicht in das Basismodell einfließen, da auch die Fragen zu diesen Merkmalen nur alle zwei Jahre gestellt werden. Bei den Jahren, in denen die befragten Personen Fragen zum Sportverhalten und dem Vorliegen chronischer Krankheiten zu beantworten hatten, handelt es sich jedoch um 2007, 2009 und 2011.

Im Laufe der Regressionsanalysen stellten sich beim Basismodell die übrigen 13 Prädiktoren als nicht geeignet heraus, um die Hospitalisierungswahrscheinlichkeit einer Person abzuschätzen. Dies ist auf verschiedene Tatsachen zurückzuführen. Einige wiesen im Rahmen der Prüfung der Merkmalsvariablen einen zu hohen p-Wert auf. Eine ausreichende Signifikanz dieser Merkmale war also nicht gegeben, sodass sie für das Modell nicht brauchbar waren.

Dies betraf Fragen bezüglich des Arbeitslebens, der Alltagseinschränkung, einiger chronischer Krankheiten, der Ernährung sowie des Versichertenstatus. Darüber hinaus konnte das Alter, in welchem Raucher mit dem Nikotinkonsum begonnen haben, nicht berücksichtigt werden, da dies lediglich in den Jahren 2002 und 2012 abgefragt wurde.

Unter den im Basismodell berücksichtigten Prädiktoren sind auch die Merkmale „Größe“ und „Gewicht“. Da jedoch große Menschen tendenziell mehr als kleinere wiegen, stellt das Kriterium Gewicht allein keinen sinnvollen Prädiktor dar.[89] Um eine bessere Einschätzung des Gewichts unter Berücksichtigung der Körpergröße der Person zu erhalten, wurden deshalb diese beiden Merkmale zusammengefasst und durch den Body-Mass-Index ersetzt. Überdies wurde das Alter bestimmt, indem das Geburtsjahr vom Erhebungsjahr subtrahiert wurde. Die folgende Tabelle stellt eine Übersicht der berücksichtigten Merkmale und der jeweiligen Merkmalsausprägung dar:

[89] *Siehe hierzu auch die im Anhang befindlich Abb. 22 zum Zusammenhang zwischen Größe und Gewicht*

Merkmal	Ausprägung
Geschlecht	Männlich/Weiblich
BMI	12,03 bis 67,47
Gesundheitszustand	Sehr gut/Gut/Zufriedenstellend/Weniger gut/Schlecht
Alter	17 bis 97 Jahre
KH	Ja/Nein
Anzahl_KH	0 bis 20 Krankenhausaufenthalte
Anzahl_Naechte	0 bis 280 Tage
Raucher	Ja/Nein
Bier	Regelmäßig/Ab und zu/Selten/Nie
Wein	Regelmäßig/Ab und zu/Selten/Nie
Spirituosen	Regelmäßig/Ab und zu/Selten/Nie
Gesundheitsbedenken	Große Sorgen/Einige Sorgen/Keine Sorgen

Tabelle 2: Prädiktoren des Basismodells und deren Ausprägungen

Wie man erkennen kann, sind einige der Merkmale metrisch, andere kategorial. Innerhalb des Datensatzes sind sie jeweils als Werte dargestellt, sodass sie in die Regression einfließen können. So werden etwa dichotome Merkmale mit die Ausprägung 0 für „nein“ bzw. 1 für „ja“ dargestellt und die Häufigkeit von Alkoholkonsum als Zahlenfolge von 1 bis 4.

Es fällt außerdem auf, dass einige Merkmale eine sehr große Spanne aufzeigen. Bei der Person, die einen BMI von nur 12 aufweist, handelt es sich beispielsweise um eine 37-jährige Frau, welche bei einer Körpergröße von 173 cm und einem Gewicht von nur 36 kg offenbar unter extremem Untergewicht leidet. Sie war im Jahr 2005 dreimal im Krankenhaus und verbrachte dort 150 Tage.

Den BMI von über 67 hatte ein 68-jähriger Mann, der bei 170 cm Körpergröße ein Gewicht von 195 kg hatte. Überraschend ist an dieser Stelle, dass er trotz dieser massiven Adipositas, welche bereits bei einem BMI von 30 beginnt,[90] im Jahr 2005 kein Krankenhaus aufsuchen musste. Außerdem schätzte er seinen Gesundheitszustand als „gut“ ein.

Die Merkmale der Tabelle 2 wurden nach dem Selektieren dahingehend gefiltert, dass nur gültige Antworten in dem Teildatensatz verbleiben. Diese Prädiktoren wurden dann in einem nächsten Schritt verwendet, um, ähnlich wie bei den einfachen Regressionsmodellen, ein Modell aufzustellen. Dabei bildet die Frage nach einem Krankenhausaufenthalt im Folgejahr den Y-Wert, also die abhängige Variable.

[90] Vgl. World Health Organization (2017), S. IV

Neben den relevanten Merkmalen, die für die 17.755 befragten Personen vorliegen, berücksichtigt das Modell überdies Interaktionen zwischen bestimmten Regressoren. Diese sind dann sinnvoll, wenn eine Linearkombination der Regressoren nicht ausreicht, um die die abhängige Variable zu erklären. Das bedeutet, dass die Effekte zweier Prädiktoren nicht einfach überlagern, sondern eine Beziehung zwischen ihnen besteht.[91]

Ein Beispiel hierfür stellt die Beziehung zwischen dem Alter und dem BMI dar, da ältere Menschen tendenziell einen höheren BMI aufweisen als jüngere (vgl. auch Abb. 6). Beim Basismodell werden neben der Interaktion zwischen Alter und BMI auch die Beziehung zwischen dem Alter und dem Raucherverhalten, der Anzahl der Krankenhausaufenthalte („Anzahl_KH“) und der Anzahl der dort verbrachten Nächte („Anzahl_Naechte“) sowie zwischen Gesundheitszustand und –bedenken berücksichtigt.

Die beiden Regressionsmodelle werden in R erzeugt. Dabei handelt es sich einerseits um ein Linear Model (LM), andererseits um ein Generalized Linear Model (GLM), welches die logistische Regression darstellt. Hieraus resultieren jeweils die einzelnen Regressionskoeffizienten. Diese sind tabellarisch aufgeführt:

Merkmal	LinReg		LogReg	
	Koeffizient	p-Wert	Koeffizient	p-Wert
Intercept	0,203446	0,000658	-1,076769	0,082011
KH	0,025333	0,377760	0,436921	0,063949
Geschlecht	0,004822	0,379028	0,053830	0,329286
Alter	-0,004351	0,000013	-0,029435	0,002950
Gesundheitszustand	0,057763	0,000000	0,258870	0,000198
Gesundheitsbedenken	0,019524	0,052777	-0,199336	0,065154
Raucher	-0,051456	0,001846	-0,540236	0,003093
Bier	0,005393	0,059775	0,055003	0,061345
Wein	0,002135	0,476883	0,006912	0,820275
Spirituosen	-0,003086	0,418719	-0,032029	0,415648
BMI	-0,006683	0,000042	-0,045593	0,009113
Anzahl_KH	0,018739	0,019056	0,073002	0,251801
Anzahl_Naechte	0,000232	0,706324	-0,004254	0,437924
Anzahl_KH:Anzahl_Naechte	0,000369	0,022934	0,003459	0,091439
Alter:BMI	0,000160	0,000001	0,001085	0,000510
Alter:Raucher	0,001053	0,002471	0,009933	0,005661
Gesundheitszustand:Gesundheitsbedenken	-0,017810	0,000000	-0,026009	0,463898

Tabelle 3: Koeffizienten und p-Werte des Basismodells, LinReg u. LogReg

Es sind sowohl die Regressionskoeffizienten β_k als auch die p-Werte dargestellt, welche die Signifikanz widerspiegeln. „Intercept“ stellt die Regressionskonstante β_0

[91] Vgl. Schlittgen, R. (2013), S. 24

dar. Die unteren vier Zeilen stellen die berücksichtigten Interaktionen der jeweiligen Prädiktoren dar. Bei den Regressionskoeffizienten der logistischen Regression muss beachtet werden, dass die Linearkombinationen aus den β-Werten und den jeweiligen X_{ik}, also die Logits, noch in die logistische Funktion

$$P(Y = 1) = \frac{e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{-z_i}}$$

eingesetzt werden müssen, um die Prädiktion für das Jahr 2007 zu erhalten.

Der nächste Schritt der Regressionsanalyse ist die Interpretation der Koeffizienten. Diese fällt bei der logistischen Regression aufgrund der logarithmierten Odds entsprechend schwerer. Es sind jedoch bei beiden Regressionstypen ähnliche Trends zu erkennen.[92] Das Merkmal „KH“ beantwortet die Frage danach, ob im Vorjahr ein Krankenhausaufenthalt stattgefunden hat. Dieses Merkmal hat entsprechend die Ausprägungen „ja“ und „nein“ bzw. 1 und 0. Dies bedeutet, dass laut beiden Modellen ein Krankenhausaufenthalt im Vorjahr die Wahrscheinlichkeit eines Aufenthaltes im Folgejahr erhöht. Es fällt dabei auf, dass der Koeffizient bei der logistischen Regression höher ist als bei der linearen.[93]

Darüber hinaus ist auffällig, dass die Koeffizienten zum Merkmal „Raucher“ jeweils negativ sind, die Hospitalisierungswahrscheinlichkeit offenbar senken. Diese überraschende Erkenntnis deckt sich mit den Beobachtungen zur Rauchergesundheit, welche bei der Darstellung ausgewählter Merkmale gemacht worden sind. Rauchen scheint das Risiko eines künftigen Krankenhausaufenthaltes für dieses Kollektiv offenbar nicht zu erhöhen (vgl. auch Anhang, Abb. 17).

Ebenso scheinen Trinkgewohnheiten und Geschlecht keinen allzu großen Einfluss darauf zu haben, ob einer Person eine stationäre Behandlung bevorsteht. Der aktuelle Gesundheitszustand hingegen, hat eine positive Wirkung auf die Wahrscheinlichkeit, da die β-Werte für dieses Merkmals positiv sind. Die Ausprägung des Gesundheitszustandes nimmt Werte von 1 („sehr gut“) bis 5 („schlecht“) an. Je höher das X_i also ist, desto höher ist laut den Regressionen auch die Wahrscheinlichkeit eines Krankenhausbesuchs in 2007.

Bei manchen Merkmalen, etwa der Anzahl der Nächte, die im Vorjahr im Krankenhaus verbracht wurden, haben die Koeffizienten beider Modelle unterschiedliche

[92] Vgl. Backhaus, K. et al. (2016), S. 313
[93] *Hierbei gilt wiederum zu beachten, dass die Logits noch in die log. Funktion einzusetzen sind.*

Vorzeichen. Dass bei der logistischen Regression ein negativer Wert als Koeffizient geschätzt wird, erscheint zunächst unrealistisch. Da jedoch die Werte in beiden Fällen nahezu null sind, hat dies nur dann einen erkennbaren Einfluss auf die Prädiktion, wenn die Anzahl der Nächte im Krankenhaus sehr hoch war.

Dennoch weisen Befragte mit einer sehr hohen Anzahl an Krankenhausnächten häufig hohe Prädiktionen auf. Dies ist unter anderem darauf zurückzuführen, dass der Gesundheitszustand, welcher einen hohen Einfluss hat, in der Regel als „Schlecht“ eingestuft wurde. Überdies ist hier die Interaktion zwischen der Anzahl der Aufenthalte zu der Anzahl der Nächte zu beachten.

Wie zu Beginn dieses Kapitels erwähnt, wird das Bestimmtheitsmaß bzw. das Pseudo-Bestimmtheitsmaß im multiplen Fall auf gleiche Weise wie bei der einfachen Regressionsanalyse ermittelt. Das bedeutet für die multiple lineare Regression, dass das Bestimmtheitsmaß wie folgt zu ermitteln ist:

$$R^2 = \frac{\sum_{i=1}^{I}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} = \frac{99{,}57}{1852{,}38} \approx 0{,}053752 \mathrel{\widehat{=}} 5{,}3752\%$$

Es ist offenbar deutlich höher als das R^2 der einfachen linearen Regression, welches rund 1,9542% beträgt. Ähnliches gilt für das Pseudo-Bestimmtheitsmaß der logistischen Regression, welches mithilfe des Nullmodells berechnet wird:

$$R^2{}_{McF} = 1 - \left(\frac{-6041}{-6456}\right) = 0{,}064157 \mathrel{\widehat{=}} 6{,}4157\%$$

Da nun der Unterschied zwischen vollständigem Modell und Nullmodell größer ist, ist folglich das $R^2{}_{McF}$ größer. Verglichen mit dem Wert, der unter Zugrundelegung des Alters als einzigen Prädiktoren ermittelt worden ist, stellen die 6,416% ebenfalls eine deutliche Steigerung dar. Außerdem gilt auch multiplen Fall, dass das logistische Modell brauchbarere Ergebnisse liefert als die lineare Regression.

Im Rahmen der Überprüfung der Modellgüte ist an dieser Stelle die Area under the curve anzuführen. Wie das Bestimmtheitsmaß auch, so ist die AUC eine standardisierte Größe, die anzeigt, wie gut ein Modell angepasst ist bzw. wie gut es dazu geeignet ist, brauchbare Prädiktionen zu liefern. Wie der Name vermuten lässt, handelt es sich bei diesem Gütemaß um eine Fläche unter einer Kurve. Wie diese Kurve wiederum definiert ist, wird im Folgenden anhand der logistischen Regression erläutert.

Das Ziel der logistischen Regression ist es, vorherzusagen, ob eine Person ins Krankenhaus kommt. Die Werte, die sich durch Einsetzen der Koeffizienten und Prädiktoren in die logistische Funktion ergeben, können als Wahrscheinlichkeit für einen Krankenhausaufenthalt interpretiert werden. Soll eine Aussage darüber getroffen werden, ab welcher Wahrscheinlichkeit man beispielsweise auf die Person zugehen möchte, um Präventivmaßnahmen einzuleiten, so muss eine Schwelle festgelegt werden. Alle Personen, die diese Schwelle dann überschreiten, erhalten vom Versicherer beispielsweise ein Informationsschreiben oder Ähnliches.

Im vorangegangen Kapitel wurde bereits festgestellt, dass ein Aufrunden ab einer 50%-Wahrscheinlichkeit oft nicht zielführend ist. Zwar würde es beim Basismodell im Gegensatz zum einfachen Modell Personen geben, für welche $P(Y = 1) \geq 0{,}5$ gilt. Von den 17.755 Befragten des untersuchten Kollektivs erfüllen dies jedoch gerade einmal 43 Personen, was lediglich rund 0,24% ausmacht. Es stellt sich also die Frage, welcher Schwellenwert am sinnvollsten wäre.

Wie zuvor erwähnt, ist diese Frage auch von der Risikoaffinität des Entscheiders und der Natur der vorherzusagenden Variable abhängig. Man könnte die Schwelle, auch „Cutoff“ genannt, niedrig ansetzen, könnte also etwa festlegen, dass ab einer Prädiktion von 5% von einem Krankenhausaufenthalt im Folgejahr ausgegangen wird. Dadurch würde die Zahl derjenigen, für welche eine Hospitalisierung vorausgesagt wurde und diese auch tatsächlich eingetreten ist, relativ hoch sein. Dies ist grundsätzlich erfreulich und spricht zunächst für eine gute Wahl der Schwelle.

Die Anzahl derjenigen, bei denen jedoch ein Aufenthalt falsch vorausgesagt wurde, weil sie im Folgejahr tatsächlich aber nicht im Krankenhaus waren, wäre entsprechend auch sehr hoch. Dies ist wiederum nicht erwünscht. Das Festlegen eines sehr niedrigen Schwellenwertes kann dann sinnvoll sein, wenn man möglichst wenige Prädiktionen erhalten möchte, bei denen falsch „tritt nicht ein“ vorhergesagt wurde, obwohl der entsprechende Fall tatsächlich eingetreten ist. Ein Beispiel hierfür stellt etwa eine HIV-Diagnose dar. Hier wird lieber einmal zu viel trotz Nichtvorliegen einer Krankheit ein positives Ergebnis vorausgesagt als einmal zu wenig trotz Vorliegen einer Krankheit eine negative Diagnose zu stellen.

Im Falle der Krankenhauswahrscheinlichkeit stellt sich also die Frage, wo man die Schwelle am besten setzt, damit möglichst wenig falsche Prädiktionen gemacht werden. Hierzu sind zunächst die Häufigkeiten der tatsächlichen Y-Werte zu zählen. Diese sind erst in Klassen eingeteilt worden, indem sie auf drei Nachkommastellen gerundet und dann nach Prädiktion in Gruppen zusammengefasst worden sind.

Dies wurde sowohl für diejenigen durchgeführt, die im Folgejahr tatsächlich im Krankenhaus waren, als auch für jene, die kein Krankenhaus aufsuchen mussten. Für die logistische Regression mit den Daten des Basismodells ergibt sich entsprechend die folgende Häufigkeitsverteilung der Krankenhausaufenthalte, jeweils gegliedert nach Klassen:

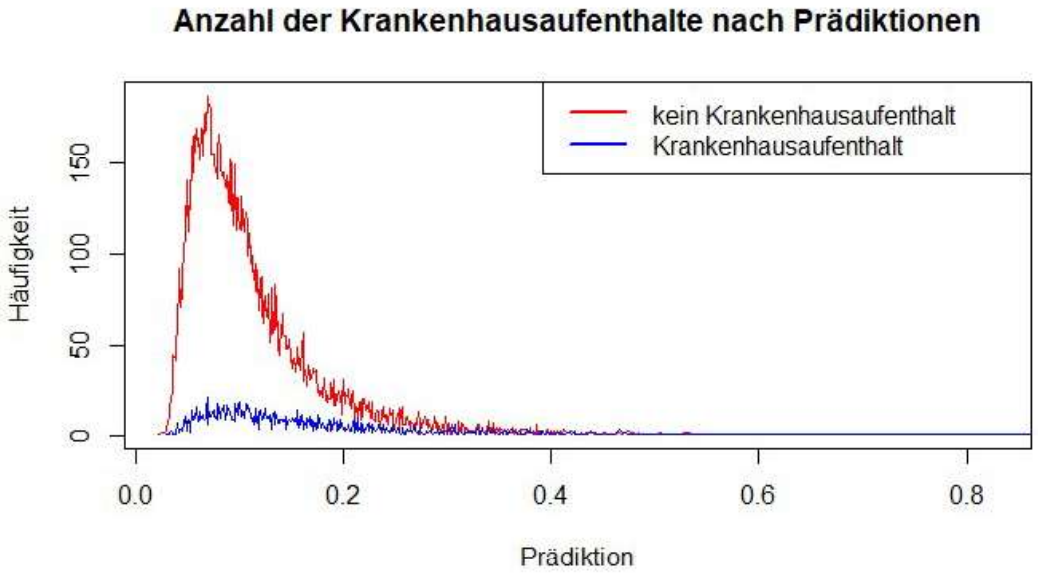

Abbildung 10: Häufigkeitsverteilung der Krankenhausaufenthalte, LogReg

Da der Anteil derjenigen, die in 2007 tatsächlich im Krankenhaus waren, lediglich rund 11,83% beträgt, verläuft die blaue Kurve deutlich unterhalb der roten Kurve. Es fällt jedoch auch auf, dass die blaue Kurve weniger stark abfällt als die rote, sodass im Bereich der hohen Prädiktionen der Anteil derjenigen, die in 2007 nicht im Krankenhaus waren, nicht mehr überwiegt.

In einem nächsten Schritt bedarf es eines festen Cutoffs. Dieser liege beispielhaft bei 10%. Dies bedeutet, alle Wahrscheinlichkeiten, die links von dieser vertikalen Linie liegen, werden als Nichteintritt bewertet, alle anderen als Eintritt. Entsprechend sind alle Häufigkeiten der roten Kurve, deren Prädiktion kleiner als 10% war, als richtige Prädiktion anzusehen. Gleiches gilt für alle Häufigkeiten der blauen Kurve, für die eine Wahrscheinlichkeit größer oder gleich dem Cutoff-Wert geschätzt worden ist. Dies lässt sich auch tabellarisch darstellen:

Allgemein		Beobachtung		**Total**
		Y=1	Y=0	
Prädiktion	Y=1	TP	FP	**TP+FP**
	Y=0	FN	TN	**FN+TN**
Total		**TP+FN**	**FP+TN**	**TP+FP+TN+FN**

Tabelle 4: Allgemeine Darstellung der Ergebnisse

Die vier Felder stellen hier, abhängig von der Höhe des Schwellenwerts, absolute Häufigkeiten dar. Dabei handelt es sich um „true positives“ (TP), „false positives“ (FP), „false negatives“ (FN) sowie um „true negatives“ (TN).[94] Dann stellt beispielsweise die Summe aus TP und FN entsprechend die Anzahl derjenigen dar, bei denen ein Eintritt beobachtet werden konnte. Daraufhin werden die sog. Sensitivität (Sens) und die Spezifizität (Spec) jeweils wie folgt berechnet:

$$Sens = \frac{TP}{TP + FN}$$

$$Spec = \frac{TN}{FP + TN}$$

Die Sensitivität stellt somit den Anteil der korrekt als positiv vorhergesagten Werte an der Gesamtheit aller positiven Beobachtungen dar. Die Spezifizität stellt auf der anderen Seite den Anteil der korrekten negativen Prädiktionen an allen negativen Beobachtungen dar. Für das Beispiel der logistischen Regression unter Berücksichtigung der Daten des Basismodells ergibt sich folglich:

LogReg, Basismodell		Beobachtung		**Total**
		Y=1	Y=0	
Prädiktion	Y=1	1.449	6.873	**8.322**
	Y=0	652	8.781	**9.433**
Total		**2.101**	**15.654**	**17.755**

Tabelle 5: Darstellung der Ergebnisse der LogReg bei 10% Cutoff

Auf Basis dessen können nun Sensitivität und Spezifizität bei einem Schwellenwert von 10% ermittelt werden. Diese sind lediglich durch Einsetzen in die bekannten Gleichungen zu berechnen. Folglich beträgt die Sensitivität rund 0,6897 und die Spezifizität etwa 0,5609. Um ein möglichst gutes Verhältnis von richtigen zu falschen Prädiktionen, also eine gute Modellanpassung, zu erhalten, ist ein denkbares Vorgehen, die Schwelle so festzulegen, dass sowohl die Sensitivität als auch die Spezifizität möglichst hoch sein sollen.

[94] Vgl. Larner, A. J. (2015), S. 168

Dieses Ziel wird bei der Ermittlung des sog. Youden-Index verfolgt. Dieser Index ist eine Kennzahl, welche Sensitivität und Spezifizität zu einem einzelnen Wert kombiniert. Dessen Maximierung gilt im Allgemeinen als sinnvoll erachtete Größe für die Festlegung eines Schwellenwerts. Der Youden-Index J wird berechnet, indem von der Summe aus Sens und Spec der Wert 1 subtrahiert wird:[95]

$$J = Sens + Spec - 1$$

Im vorliegenden Fall eines Schwellenwerts von 10% beträgt der Youden-Index demnach $J = 0{,}6897 + 0{,}5609 - 1 = 0{,}2506$. Würde man ab einer Prädiktion von rund 11,89% von einem Krankenhausaufenthalt in 2007 ausgehen, so würde der Youden-Index einen Wert von etwa 0,2702 annehmen. Dies scheint also ein besserer Schwellenwert als 10% zu sein. Jedoch ist auch ein Cutoff mit einem maximalen Youden-Index nicht immer die perfekte Größe, welche für alle Modelle gilt.

Deshalb wird eine sog. Receiver-Operating-Characteristic-Kurve (ROC-Kurve) gebildet. Diese entsteht durch Festlegen des Schwellenwerts auf alle Werte zwischen 0 und 1. Dies wird anhand der logistischen Regression beispielhaft dargestellt, indem einige Schwellenwerte und die entsprechenden Sensitivitäten und Spezifizitäten ermittelt werden. Die Werte von Sens und Spec in Abhängigkeit der jeweiligen Cutoff-Werte finden sich in folgender Tabelle wieder:

Cutoff	Sens	Spec	1-Spec	Youden-Ind.
0%	1,0000	0,0000	1,0000	0,0000
2,5%	1,0000	0,0001	0,9999	0,0001
5%	0,9738	0,0810	0,9190	0,0548
7,5%	0,8453	0,3370	0,6630	0,1824
10%	0,6897	0,5609	0,4391	0,2506
12,5%	0,5502	0,7135	0,2865	0,2637
15%	0,4441	0,8088	0,1912	0,2529
17,5%	0,3508	0,8686	0,1314	0,2194
20%	0,2751	0,9050	0,0950	0,1801
22,5%	0,2213	0,9325	0,0675	0,1539
25%	0,1766	0,9509	0,0491	0,1275
27,5%	0,1409	0,9641	0,0359	0,1050
30%	0,1195	0,9737	0,0263	0,0932
…	…	…	…	…
100%	0,0000	1,0000	0,0000	0,0000

Tabelle 6: Sens, Spec und Youden-Index auf Basis verschiedener Cutoffs

[95] Vgl. Larner, A. J. (2015), S. 168

Zwar sind die Schritte hier nicht sehr klein, damit die Tabelle ihre Übersichtlichkeit nicht verliert. Dennoch ist auch bei 2,5%-Schritten ein gewisser Verlauf erkennbar. Der Cutoff, welcher zum höchsten Youden-Index führt, liegt offenbar zwischen 10% und 15%. Die Spalte „1-Spec“ ist deshalb Teil der Tabelle, weil dies bei der Darstellung der ROC-Kurve die Werte der X-Achse definiert. Die Werte der Spalte „Sens“ stellen die Y-Werte der Kurve dar.

Zeichnet man die X- und Y-Werte also in ein Koordinatensystem, so ergibt sich daraus allmählich die ROC-Kurve. Es gilt dabei, dass je mehr Schwellen angesetzt worden sind, desto genauer die ROC-Kurve ist. Die folgende Abbildung zeigt die Andeutung derjenigen ROC-Kurve, welche sich aus den Werten der obigen Tabelle ergibt:

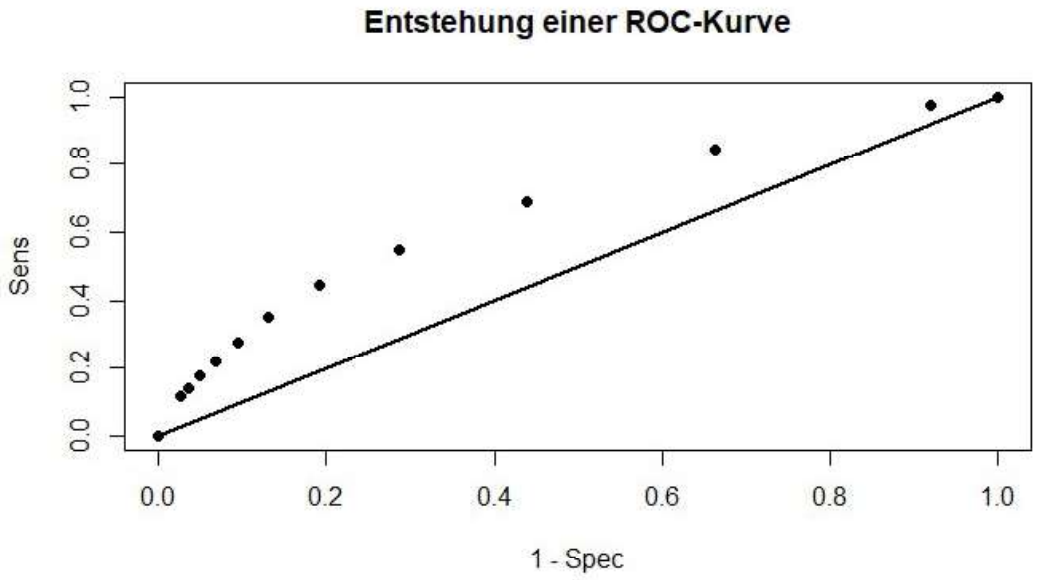

Abbildung 11: Entstehung einer ROC-Kurve

Es fällt auf, dass die Abstände der Punkte nicht immer gleich sind, obwohl die Schwelle in 2,5%-Schritten verändert wurde. Dies ist auf die spezielle Verteilung der Häufigkeiten der Prädiktionen des logistischen Regressionsmodells mit den Basisdaten zurückzuführen. Die Diagonale, welche in einem 45°-Winkel verläuft, stellt die schlechteste Anpassung dar. Eine ROC-Kurve, die genau auf dieser Linie verlaufen würde, würde also auf ein völlig unbrauchbares Modell hindeuten.

Im Rahmen der Bewertung der Modellgüte wird nun deutlich, dass es sich bei der Area under the curve und die Fläche unter der ROC-Kurve handelt. Neben dem Bestimmtheitsmaß bzw. dem Pseudo-Bestimmtheitsmaß im Fall der logistischen

Regression stellt die AUC einen Standard-Wert dar, welcher es erlaubt, Modelle miteinander zu vergleichen. Die Berechnung der Fläche kann mithilfe verschiedener Methoden der Integralrechnung geschehen.[96] Da die AUC jedoch stets mithilfe von R berechnet wird, wird auf die genaue Methodik der Berechnung in dieser Arbeit nicht weiter eingegangen.

Der Wertebereich der AUC als Gütemaß nimmt Werte zwischen 0,5 und 1 an. Dabei gilt 1 als perfekte Anpassung. Liegt die Kurve jedoch genau auf der o.g. Diagonalen, so beträgt die darunter liegende Fläche folglich nur 0,5. Berechnet man die AUC unter den ROC-Kurven der bislang vorliegenden Modelle, so erhält man bei Durchführung der linearen Regression im einfachen Fall ein AUC von 0,6231 und im multiplen Fall 0,6784. Die logistische Regressionsanalyse führt ebenfalls zu 0,6231 im einfachen Fall bzw. zu 0,6795 im multiplen Fall.

Diese Ergebnisse unterscheiden sich zwar von den Bestimmtheitsmaßen. Dennoch sind auch hier auch hier Tendenzen zu erkennen. Einerseits, dass multiple Modelle aussagekräftiger sind, andererseits, dass die logistische Regression zu besseren Ergebnissen als die lineare Regression führt. Dabei gilt, wie für den einfachen Fall auch, dass die Prädiktionen aus der linearen Regression, im Gegensatz zur logistischen Regression, auch Werte < 0 und > 1 annehmen können, was ein klarer Vorteil des logistischen Ansatzes ist.

Nachdem das Training auf Basis des Jahres 2006 abgeschlossen ist, wird das Modell nun mit seinen jeweiligen Regressionskoeffizienten auf das Jahr 2007 angewendet. Hierbei wird getestet, wie brauchbar das Modell für ein anderes Jahr als das Basisjahr ist. Dabei bleiben die β-Werte unverändert. Sie werden jedoch auf neue X-Werte angewendet. Die Personen, die dies betrifft, müssen hierbei nicht dieselben aus 2006 sein.

Für das Jahr 2007 lagen die Angaben zu Gewicht und Größe sowie dem Raucher- und Alkoholverhalten nicht vor. Aufgrund dessen wurde eine Interpolation dieser Merkmale vorgenommen, indem der Durchschnitt der Merkmale aus 2006 und 2008 verwendet wurde. Lediglich bei dem Merkmal „Raucher“ wurde davon abgesehen, weil es sich um eine dichotomes Merkmal handelt, dessen Ausprägung nur „ja“ und „nein“ ist. Stattdessen wurde angenommen, dass das Raucherverhalten sich innerhalb eines Jahres nicht verändert hat. Entsprechend wurden hier die Werte aus 2006 fortgeschrieben.

---

[96] Vgl. Centor, R. M. / Schwartz, J. S. (1985), S. 151 ff.

Wendet man nun das Modell aus dem Jahr 2006 auf die Daten des Jahres 2007 an, so ergeben sich wieder neue Prädiktionen der Krankenhauswahrscheinlichkeit im Folgejahr, in diesem Fall also für 2008. Diese Prädiktionen können wiederum unter Festlegung einer bestimmten Schwelle dazu verwendet werden, die Sensitivität und die Spezifizität zu berechnen. Auch hier wird die ROC-Kurve gebildet, indem alle Schwellenwerte von 0 bis 1 berücksichtigt werden.

Der Verlauf der ROC-Kurve für das Jahr 2007 ist dem des Jahres 2006 sehr ähnlich. Die folgende Abbildung zeigt die ROC-Kurve, welche grundsätzlich durch die beschriebene Methodik zu bilden ist. Die vorliegende Kurve wurde mithilfe des R-Pakets „pROC“ gebildet.[97] Es handelt sich hierbei um die Kurve für die logistische Regression:

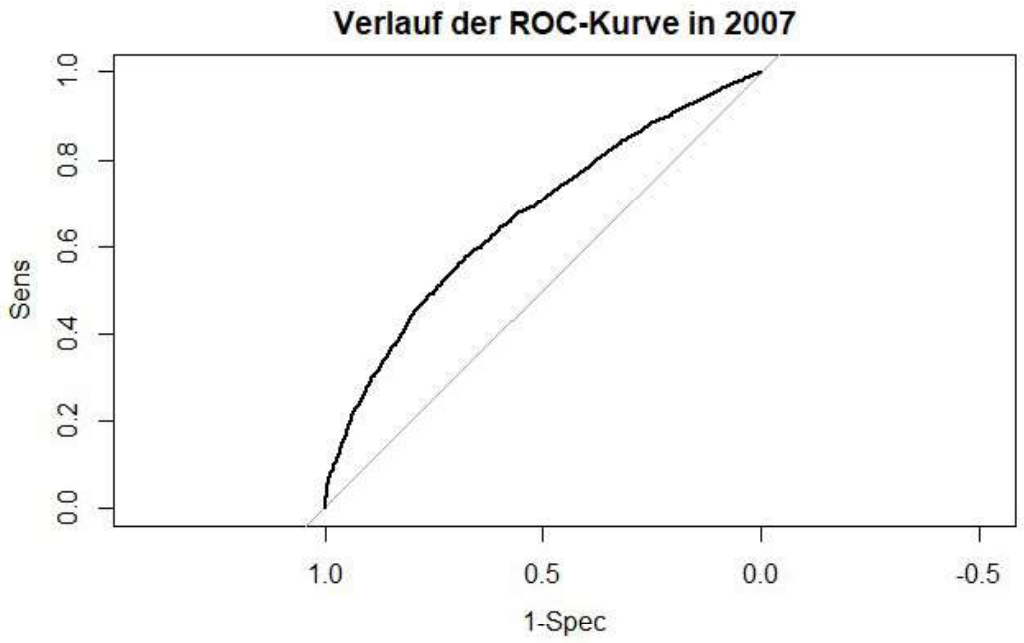

Abbildung 12: Verlauf der ROC-Kurve in 2007, LogReg

Die Kurve entsteht nach der erläuterten Methodik, also durch das Anwenden zahlreicher Cutoff-Werte. Die Fläche, welche die Kurve mit der X-Achse einschließt, stellt dann die für die Prüfung der Modellgüte relevante Area under the curve dar. Darüber hinaus wird das bekannte Gütemaß, das Bestimmtheitsmaß, für die lineare Regression ermittelt. Eine Berechnung des Pseudo-Bestimmtheitsmaß nach McFadden ist für Folgejahre jedoch nicht möglich.

Der Grund hierfür liegt in der Formel zur Berechnung des R^2_{McF}. Hier wird der Quotient aus der Summe der Log-Likelihoods des vollständigen Modells und der Summe

[97] Robin, X. et al. (2011)

der Log-Likelihoods des Nullmodells gebildet. LL_0 legt dabei die Daten aus 2007 zugrunde. Bei der Berechnung von LL_v hingegen, werden die Koeffizienten des Modells betrachtet. Da diese jedoch auf Basis des Jahres 2006 berechnet worden sind, werden entsprechend nur die Daten dieses Jahres berücksichtigt, sodass der Dividend bzw. Zähler für alle Folgejahre den gleichen Wert annimmt. Daraus resultieren bei der Berechnung des $R^2{}_{McF}$ unbrauchbare Ergebnisse der Modellgüte.

Aufgrund der Tatsache, dass auch überprüft werden soll, wie zeitkonsistent das Modell ist, werden die Koeffizienten aus 2006 auch auf die unabhängigen Variablen des Jahres 2010 angewendet. Für dieses Jahr liegen die benötigten Merkmale zur Ermittlung der Prädiktion für 2011 vor, sodass eine Interpolation hier nicht notwendig ist. Entsprechend sind auch für dieses Jahr das Bestimmtheitsmaß der linearen Regression sowie die AUC beider Regressionsmodelle leicht zu ermitteln. Lediglich das Pseudo-Bestimmtheitsmaß nach McFadden ist auch für 2010 nicht vorhanden. Alle Ergebnisse, die auf Basis der Koeffizienten des Basismodells von 2006 ermittelt worden sind, sind in der folgenden Tabelle dargestellt:

Basismodell *(2006)*		**Gütemaß**	
		R^2	AUC
LinReg	Nur Alter	1,954%	62,31%
	Training 2006 (2006)	5,375%	67,84%
	Testing 2007	4,887%	66,23%
	Testing 2010	5,232%	67,74%
LogReg	Nur Alter	2,705%	62,31%
	Training 2006 (2006)	6,416%	67,95%
	Testing 2007	-	66,48%
	Testing 2010	-	67,87%

Tabelle 7: Gütemaße des Basismodells

Diese Übersicht verdeutlicht noch einmal, dass das einfache Modell, das nur das Alter als Regressor verwendet, wenig Aussagekraft besitzt. Außerdem lassen sich in der Tabelle leicht die Ergebnisse von (multipler) linearer Regression und logistischer Regression vergleichen. Überdies sind anhand der Ergebnisse der Jahre 2007 und 2010 Aussagen zur zeitlichen Konsistenz der Regressionsmodelle möglich.

Es ist zu beobachten, dass die Werte für das Jahr 2007 leicht gesunken sind. Auf der anderen Seite sind sie im Jahr 2010 wieder gestiegen, sodass sogar nahezu die Gütequalität des Trainings-Jahres erreicht wird. Dies spricht für eine stabile Modellanpassung. Dennoch lässt sich auch sagen, dass sowohl die Bestimmtheitsmaße als auch Verläufe der ROC-Kurven und die daraus resultierenden AUC keine allzu

hohen Werte erreichen. Aussagen dazu, dass Personen mit bestimmten Merkmalsausprägungen tendenziell eher gefährdet sind, im Folgejahr ein Krankenhaus aufsuchen zu müssen, lassen sich trotzdem machen.

3.3.3. Alternativmodell

Aufgrund der Tatsache, dass die Merkmale zu Sport und chronischen Krankheiten sich nicht mit denen des Basismodells überschneiden, wird im Folgenden ein alternatives Modell vorgestellt. Dieses berücksichtigt eben jene Daten zur sportlichen Aktivität und zum Vorliegen chronischer Krankheiten, die im vorangegangenem Kapitel keine Anwendung fanden. Auch mit diesen Daten wird eine lineare und eine logistische Regressionsanalyse durchgeführt.

Anschließend wird auch für das Alternativmodell eine Interpretation der Regressionskoeffizienten vorgenommen und die Ergebnisse der Analysen in Form von Gütemaßen dargestellt. Wie im Kapitel zum Basismodell auch, werden in diesem Kapitel zunächst die berücksichtigten Merkmale tabellarisch aufgeführt und erläutert. Das Jahr, auf dessen Basis die Regressionsmodelle aufgestellt werden, ist beim Alternativmodell 2009. Die folgende Tabelle zeigt eine Übersicht der Merkmale:

Merkmal	**Ausprägung**
Geschlecht	Männlich/Weiblich
Gesundheitszustand	Sehr gut/Gut/Zufriedenstellend/Weniger gut/Schlecht
Alter	18 bis 99
Diabetes	Ja/Nein
Herzerkrankung	Ja/Nein
Krebs	Ja/Nein
Schlaganfall	Ja/Nein
Hypertonie	Ja/Nein
Depression	Ja/Nein
Sonstige_Krankheit	Ja/Nein
Keine_Krankheit	Ja/Nein
KH	Ja/Nein
Anzahl_KH	0 bis 28
Anzahl_Naechte	0 bis 290
Gesundheitsbedenken	Große Sorgen/Einige Sorgen/Keine Sorgen
Sport	Jede Woche/Jeden Monat/Selten/Nie

Tabelle 8: Prädiktoren des Alternativmodells und deren Ausprägungen

Insgesamt werden beim Alternativmodell also 16 Prädiktoren betrachtet. Diese dienen dazu, die abhängige Variable Y, also die Frage nach einem Krankenhausaufenthalt im Jahr 2010, abzuschätzen. Die oben genannten chronischen Krankheiten wurden zusammenfassend als diese bezeichnet, da Krankheiten wie beispielsweise Diabetes auch im Folgejahr noch bestehen. Überdies beinhaltet dieser Teil des SOEP-Datensatzes die Merkmale „Sonstige_Krankheit“ und „Keine_Krankheit“. Die Anzahl der Personen des Jahres 2009 beträgt nach der Selektion und Filterung der Merkmale 14.908.

Auch in diesem Datensatz wurde das Alter ermittelt, indem die Differenz zwischen Erhebungsjahr und Geburtsjahr der befragten Person gebildet wurde. Außerdem findet, wie beim Basismodell, auch bei den Regressionsmodellen des Alternativmodells eine Berücksichtigung von Interaktionen statt. Dies betrifft die Beziehung zwischen der Anzahl der Krankenhausaufenthalte und der Anzahl der Nächte, zwischen dem Gesundheitszustand und –bedenken sowie zwischen den sportlichen Aktivitäten und einer Hypertonie.

Die Schätzung der einzelnen Regressionskonstanten und –koeffizienten wird mithilfe der in vorangegangenen Kapiteln erläuterten Schätzmethoden durchgeführt. Die β- und p-Werte, die sich aus diesem Vorgehen ergeben haben, sind in der folgenden Tabelle enthalten:

Merkmal	LinReg		LogReg	
	Koeffizient	p-Wert	Koeffizient	p-Wert
Intercept	0,061790	0,177171	-2,912486	6,1E-14
KH	-0,043140	0,151072	0,036067	0,869990
Geschlecht	0,021620	0,000068	0,203231	0,000064
Alter	0,001150	7,68E-10	0,011630	6,07E-11
Gesundheitszustand	0,073020	5,94E-16	0,402228	1,79E-07
Gesundheitsbedenken	0,048650	0,000039	0,141814	0,224420
Diabetes	0,015800	0,000023	0,089144	0,001390
Herzerkrankung	0,019850	1,42E-08	0,102534	0,000080
Krebs	0,011790	0,018569	0,064949	0,085080
Schlaganfall	0,017480	0,020262	0,083974	0,109380
Hypertonie	0,005659	0,282277	0,057968	0,204110
Depression	0,004235	0,338299	0,029989	0,388140
Sonstige_Krankheit	0,008327	0,003289	0,052186	0,022800
Keine_Krankheit	-3,005E-05	0,991729	-0,040525	0,126790
Anzahl_KH	0,032510	0,000103	0,138089	0,021380
Anzahl_Naechte	0,002852	0,000004	0,010945	0,010930
Sport	0,001667	0,575842	0,010624	0,668790
Anzahl_KH:Anzahl_Naechte	-0,000447	0,007193	-0,001852	0,135210
Gesundheitszustand:Gesundheitsbedenken	-0,022570	6,35E-08	-0,082163	0,029030
Sport:Hypertonie	-0,001341	0,405287	-0,019435	0,158530

Tabelle 9: Koeffizienten und p-Werte des Alternativmodells, LinReg u. LogReg

Für die logistische Regression gilt, dass die Logits, die sich aus den Linearkombinationen von Merkmalsausprägungen und Regressionskoeffizienten ergeben, in die logistische Funktion einzusetzen sind, um die jeweiligen Prädiktionen zu erhalten. Grundsätzlich ist die Wirkung der meisten Koeffizienten auf die Hospitalisierungswahrscheinlichkeit der Wirkung beim Basismodell relativ ähnlich. Es sind jedoch Unterschiede beim Ausmaß des Einflusses eines Prädiktors festzustellen.

Es fällt auf, dass die Regressionskonstanten beider Regressionen beim Alternativmodell geringer sind. Darüber hinaus hat beim Alternativmodell das Merkmal „KH“, welches die Frage nach einem Krankenhausaufenthalt im Vorjahr beantwortet, offenbar einen geringeren Einfluss auf den Y-Wert des Modells. Bei der linearen Regression nimmt dieses β_k sogar einen negativen Wert an. Der Gesundheitszustand, der beim Basismodell bereits eine zentrale Rolle spielt, hat einen relativ hohen Einfluss auf die Wahrscheinlichkeit für eine stationäre Behandlung im Folgejahr.

Die Regressoren, die die chronischen Krankheiten beschreiben, haben zwar alle einen positiven Einfluss auf die Krankenhauswahrscheinlichkeit. Da die Merkmalsausprägung hier jedoch nur „ja“ bzw. „nein“ ist, ist der Einfluss von dieser Krankheiten relativ gering. Das Merkmal „Keine_Krankheit“ hat, vor allem bei der multiplen linearen Regression, kaum einen Einfluss auf die Hospitalisierung. Dies gilt, ebenso wie beim Basismodell, auch für die gewählten Interaktionen. Auch Sport hat nur geringfügig von 0 abweichende β-Werte, der Einfluss dieses Regressors ist demnach auch nicht sehr groß.

Im Rahmen der Prüfung des Gesamtmodells ergeben sich beim Alternativmodell das folgende Bestimmtheitsmaß der linearen bzw. Pseudo-Bestimmtheitsmaß der logistischen Regression:

$$R^2 = \frac{113{,}44}{1718{,}50} \approx 0{,}066013 \mathrel{\widehat{=}} 6{,}6013\%$$

$$R^2{}_{McF} = 1 - \left(\frac{-5421}{-5843}\right) \approx 0{,}072232 \mathrel{\widehat{=}} 7{,}2232\%$$

Das Bestimmtheitsmaß sowohl der linearen als auch der logistischen Regression ist bei den Daten des Alternativmodells offenbar größer, die Modellanpassung also besser. Die Vermutung, dass das Alternativmodell eine höhere Güte als das Basismodell besitzt, wird ebenfalls von den Ergebnissen der Analyse der ROC-Kurven

gestützt. So erzielt die multiple lineare Regression eine AUC von rund 0,6835, die Fläche unter der ROC-Kurve der logistischen Regression beträgt sogar 0,6837.

Um eine noch bessere Vergleichbarkeit von Basis- und Alternativmodell zu erhalten, wird im Folgenden auch für das Alternativmodell eine Regressionsanalyse des Jahres 2010 durchgeführt. Dadurch kann verglichen werden, welches Modell zu besseren Prädiktionen von Krankenhauswahrscheinlichkeiten des Jahres 2011 führt. Aufgrund der Tatsache, dass die Daten des Alternativmodells nur in den Jahren 2009 und 2011 vorliegen, nicht aber in 2010, ist eine Interpolation durchzuführen. Dies geschieht auf gleiche Weise wie es bereits beim Basismodell für das Jahr 2007 gemacht worden ist, es wird also der Mittelwert der Daten aus 2009 und 2011 gebildet.

Dies betrifft unter anderem die Merkmale zu chronischen Krankheiten. Da diese jedoch dichotom sind, sie also lediglich die Merkmalsausprägungen „ja“ und „nein“ besitzen, ist hier wieder eine Interpolation nicht sinnvoll. Aus diesem Grund werden die Antworten aus 2009 für diese Merkmale fortgeschrieben. Dies ist deshalb relativ problemlos möglich, weil chronische Krankheiten in der Regel auch im Folgejahr noch bestehen. Ein scheinbar veralteter Wert aus dem Jahr 2009 ergibt demnach mehr Sinn als ein Durchschnitt von „ja“ und „nein“.

Die Prädiktionen für 2011 werden also in 2010 auf Basis der Koeffizienten aus 2009 berechnet. Durch Selektieren des SOEP-Datensatzes erhält man für 2010 eine Anzahl befragter Personen von 12.500. Auch bei dieser Testphase des Modells ist bei der logistischen Regression ein Ermitteln des Pseudo-Bestimmtheitsmaßes nicht möglich. Die Area under the curve hingegen, lässt sich schnell errechnen. Daraus ergeben sich für das Vergleichsjahr 2010 die folgenden Größen:

	Modell	AUC
LinReg	Basismodell (2006)	67,74%
	Alternativmodell (2009)	68,45%
LogReg	Basismodell (2006)	67,87%
	Alternativmodell (2009)	68,65%

Tabelle 10: AUC von Basis- und Alternativmodell in 2010, LinReg u. LogReg

Zwar sind die Koeffizienten des Basismodells etwas „älter“, weil sie auf Basis des Jahres 2006 anstelle des Jahres 2009 berechnet worden sind. Änderungen aufgrund von Effekten wie beispielsweise des medizinischen Fortschritts wirken also

stärker. Da es jedoch lediglich drei Jahre sind, die das Basismodell älter als das Alternativmodell ist, spielen solche Effekte kleine allzu große Rolle.

Vergleicht man die Werte der Modelle, so erkennt man, dass die AUC beim Alternativmodell größer sind. Dies trifft sowohl auf die lineare als auch auf die logistische Regression zu. Dabei gilt auch, dass die logistische Regression offenbar bessere Prädiktionen liefert als die als lineare, was nicht zuletzt auf die Wertebereiche der beiden Regressionstypen zurückzuführen ist.

Insgesamt lässt sich beobachten, dass die Prädiktionen beim Alternativmodell weniger extrem sind als die beim Basismodell. Das bedeutet, sie sind weiter von den Wahrscheinlichkeiten 0% und 100% entfernt. Dies lässt sich beispielsweise an der Verteilung der Prädiktionen der logistischen Regression der beiden Datensätze erkennen. Für das Jahr 2010 weist die logistische Regression, bei der die Koeffizienten aus 2006 verwendet wurden, eine minimale Voraussage einer Krankenhauswahrscheinlichkeit von rund 0,37% und eine maximale Prädiktion von etwa 99,98% auf.

Diese Werte sind sehr niedrig bzw. hoch. Eine sehr hohe geschätzte Wahrscheinlichkeit hat häufig auch einen Y-Wert von 1 zur Folge. Dennoch war gerade jene Person, für die eine fast hundertprozentige Prädiktion geliefert worden ist, im Jahr 2011 nicht im Krankenhaus. Es handelt sich dabei um einen 18-jährigen mit schlechtem Gesundheitszustand und großen Gesundheitsbedenken. Überdies hatte er im Vorjahr 13 Krankenhausaufenthalte mit 220 dort verbrachten Nächten.

Anders stellt sich die Kurve des Alternativmodells dar. Diese verläuft insgesamt etwas flacher als die des Basismodells (s. Anhang, Abb. 26 u. 27). Hierbei ergibt sich bei Anwendung der logistischen Regression im Jahr 2010 eine minimale Prädiktion von rund 3,61% und eine maximale von 89,89%. Zwar sind diese Werte auch relativ niedrig bzw. hoch, jedoch sind sie weniger gestreut als die des Basismodells. Die Kombination aus Koeffizienten und unabhängigen Merkmalen hat hierbei also eine geringere Wirkung auf die Wahrscheinlichkeit, dass die entsprechende Person in 2011 ein Krankenhaus aufsuchen muss.

4. Schlussbetrachtung und Ausblick

Nachdem in den vorangegangenen Kapiteln alle Modellierungen und Berechnungen durchgeführt worden sind, werden in diesem Kapitel die gewonnenen Erkenntnisse nochmal einmal zusammengefasst. Darüber hinaus wird untersucht, inwiefern eine statistische Analyse sozioökonomischer Daten zur Prädiktion von Krankenhausaufenthalten in der Praxis für ein Versicherungsunternehmen umsetzbar ist. Dabei werden sowohl Chancen als auch Risiken betrachtet.

Außerdem werden in diesem abschließendem Kapitel die Erkenntnisse aus der Sichtung der Sekundärliteratur mit den Ergebnissen der Regressionsanalysen verglichen. Dabei wird überprüft, ob und inwiefern es Überschneidungen und auch Unterschiede gibt und was mögliche Gründe für Differenzen sein können. Überdies wird ein kurzer Ausblick darüber gegeben, wie die Verwertung solcher Daten künftig aussehen kann.

Die Anwendung verschiedener Regressionsmodelle hat gezeigt, dass eine einfache Regression wenig Aussagekraft hat, da bei einer Krankenhauswahrscheinlichkeit offenbar deutlich mehr Faktoren eine Rolle spielen als lediglich das Alter. Dies wiederum beutetet nicht, dass ein Modell umso besser ist, je mehr unabhängige Variablen bei der Modellierung berücksichtigt werden. So können manche Merkmale die Modellgüte sogar verringern.

Aufgrund der Tatsache, dass bei einer linearen Regression auch negative Werte bzw. Y-Werte über 1 angenommen werden können, eignet sich diese Methode weniger gut zur Prädiktion von Wahrscheinlichkeiten. Da die Linearkombinationen bei der logistischen Regressionsanalyse logarithmiert werden, ist sie entsprechend besser geeignet, denn dadurch liegen die Prädiktionen für einen Krankenhausaufenthalt zwischen 0 und 1. Darüber hinaus lag beim Alternativmodell offensichtlich eine bessere Anpassung vor, wie der Abgleich der Prädiktionen mit den tatsächlichen Beobachtungen im Folgejahr zeigte.

Vergleicht man Ergebnisse der Regressionsanalysen mit denen aus der Sekundärliteratur, so ergeben sich an vielen Stellen Übereinstimmungen. Unter anderem ein hohes Alter, zu wenig Sport sowie ein schlechter Gesundheitszustand sind Indizien für einen bevorstehenden Krankenhausaufenthalt. Auch ein hoher BMI und einige chronische Krankheiten deuten, wenngleich weniger stark als erwartet, darauf hin, dass die Verfassung eines Menschen schlecht ist und er unter einem erhöhtem Risiko steht, im Folgejahr stationär behandelt werden zu müssen.

Es gibt jedoch auch einige überraschende und widersprüchliche Erkenntnisse. Dabei steht vor allem die Gesundheit von Rauchern im Fokus. Es gilt als allgemein bekannt und auch medizinisch nachgewiesen, dass Rauchen extrem gesundheitsschädigend wirken kann und die Lebenserwartung verringert. Mögliche Gründe dafür, dass der Gesundheitszustand der Raucher und die Häufigkeit von Krankenhausaufenthalten kaum von den bei den Nichtrauchern gemachten Beobachtungen abweichen, können unter anderem in der Raucherdefinition liegen.

So zählen möglicherweise auch Gelegenheitsraucher, die allerdings eine bessere Gesundheit aufweisen, in diese Kategorie. Dadurch erscheint die Gruppe der „echten“ Raucher insgesamt gesünder als sie eigentlich ist. Dies könnte auch für das Trinkverhalten einer Person gelten. Denn auch bei diesem Merkmal weichen die Ergebnisse der Regressionsanalysen von den Erkenntnissen der Medizin ab.

Einige Merkmale, beispielsweise Atemwegserkrankungen, Rückenleiden oder die Ernährungsgewohnheiten einer Person, sind gemäß Sekundärliteratur gesundheitsbeeinflussende Faktoren. Bei der Prüfung der einzelnen Modellparameter hat sich jedoch herausgestellt, dass diese Merkmale als nicht signifikant anzusehen sind. Aus diesem Grund fanden sie bei den Modellen keine Berücksichtigung. Dies gilt ebenfalls für das Kriterium des sozioökonomischen Status. Dieser gilt als einer der zentralen Faktoren, die einen massiven Einfluss auf die Gesundheit haben.

Der Grund dafür, dass der Sozialstatus nicht in den Modellen enthalten ist, liegt darin, dass er häufig durch das Bildungs- und Einkommensniveau sowie den Beruf einer Person definiert wird. Daten zum Beruf und zum Bildungsniveau lagen jedoch entweder aufgrund zeitlicher Differenzen oder aufgrund mangelnder Signifikanz nicht vor. Das Fehlen von Signifikanz trifft ebenfalls auf das ohnehin lückenhafte Merkmal „Monatsgehalt“ zu.

Insgesamt lässt sich aufgrund teilweise widersprüchlicher Ergebnisse die Vermutung aufstellen, dass eine Hospitalisierung im Folgejahr allein nicht zwingend ein aussagekräftiges Merkmal bei der Beurteilung der Gesundheit darstellt. Aus Sicht eines Versicherers lassen sich nur anhand dieses Merkmals überdies nicht die zu erwartenden Kosten, die ein VN verursachen wird, messen. Beispielsweise suchen Versicherungsnehmer je nach Krankheitserscheinung nicht gleich ein Krankenhaus auf, sondern werden ambulant behandelt. Dies würde zwar Kosten verursachen, das Merkmal der Hospitalisierung der Person wäre dadurch jedoch nicht erfüllt.

Darüber hinaus kommt es vor, dass eine Person, die grundsätzlich als gesund anzusehen ist, aus Gründen der Prävention ein Krankenhaus aufsucht. Diese Personengruppe ist mehr um ihre Gesundheit bedacht. Sie kann also ein geringeres Risiko für das VU darstellen als die Gruppe derjenigen, die zwar krank sind, ein Krankenhaus jedoch nur dann aufsuchen, wenn es bereits für leichtere Eingriffe zu spät ist. Hieraus wiederum können sehr hohe Kosten für den Versicherer entstehen.

Ein Zusammenhang zwischen der Art des Versichertenverhältnisses einer Person, also ob sie privat oder gesetzlich versichert ist, und der Häufigkeit von Krankenhausaufenthalten konnte im Rahmen der Analyse nicht erkannt werden. Die klassischen Tarifmerkmale „Alter“ und „Geschlecht“[98] spielen offenbar auch bei dem Kollektiv des SOEP-Datensatzes für die Prädiktion eine Rolle, obgleich diese weniger groß ist als erwartet. Dennoch lässt sich erkennen, dass hohe Prädiktionen vor allem bei älteren Personen zu finden sind, was wiederum eine Rechtfertigung der Alterungsrückstellungen der PKV darstellt.

Die Verwendung dieser sozioökonomischen Daten durch ein Versicherungsunternehmen brächte, je nach Ausmaß, sowohl Chancen als auch Risiken mit sich. Durch das Sammeln von aktuellen Daten der Versicherungsnehmer wüsste der Versicherer mehr über seinen Bestand. Die Informationen hierzu könnten es dem Versicherer erleichtern, Risiken besser einzuschätzen und zu klassifizieren, etwa durch bestimmte Teilkollektive.

Eine weitere Anwendungsmöglichkeit, die sich aus der Verarbeitung der Daten ergeben könnte, stellt das Einleiten von Präventivmaßnahmen durch den Versicherer dar. Da dieser auf Basis der gesammelten Daten Wahrscheinlichkeiten zu bevorstehenden Krankenhausaufenthalten berechnen könnte, bestünde die Möglichkeit, gezielt auf jene VN zuzugehen, deren Prädiktion einen bestimmten Schwellenwert überschreitet. Dadurch könnten, je nach Ursache der hohen Schätzung, verschiedene Maßnahmen zur Minderung des Risikos eingeleitet werden.

Die Arten der einzelnen Maßnahmen können verschieden sein. Zu wenig sportlicher Aktivität oder einem zu hohen BMI könnte etwa mithilfe von Vergünstigungen für ein Fitnessstudio oder Ähnlichem entgegengewirkt werden. Auch Informationsbroschüren sind ein denkbarer Ansatz. Dabei wäre jedoch zu beachten, ein gewisses Feingefühl an den Tag zu legen, damit der VN sich nicht in eine Situation gedrängt sieht, in der ihm ein Vorwurf gemacht wird bzw. er sich angegriffen fühlt.

---

[98] *Beim Geschlecht gilt es wiederum, die im Kapitel „2.2. Tarifmerkmale der privaten Krankenversicherung“ getroffenen Aussagen zum AGG zu beachten.*

Ein weniger individueller Ansatz kann sein, die auf Grundlage der statistischen Analyse als gesundschädlich befundenen Merkmale bzw. Verhaltensweisen zu vermindern, indem gewisse Belohnungen, etwa für einen gesunden Lebensstil, geboten werden. Dies wird teilweise bereits praktiziert. So werden im Rahmen von Bonusprogrammen etwa Nachweise einer Mitgliedschaft in einem Sportverein, Schutzimpfungen oder regelmäßige Kontrollen beim Zahnarzt mit kleinen Prämien honoriert.[99]

Die Umsetzung würde jedoch auch Herausforderungen mit sich bringen. So werden die Versicherungsnehmer beispielsweise nicht jährlich nach ihrem aktuellen BMI, ihrem Raucherverhalten oder ihrer sportlichen Aktivität befragt. Die Datenerhebung wäre folglich für manche Merkmale leicht, für andere wiederum problematischer. Durch Fitness-Tracker könnte das Sportverhalten einer Person auf freiwilliger Basis analysiert werden. Daten zu vielen relevanten Merkmalen könnten überdies durch Auswertung der Leistungsabrechnungen des Krankenversicherers erhoben werden.

Allerdings ist ein bereits versicherter Versicherungsnehmer zum Beispiel nicht verpflichtet, dem Versicherungsunternehmen mitzuteilen, dass er inzwischen Raucher ist. Da ihm in der Regel bekannt sein wird, dass dies kein prämienminderndes Merkmal darstellt, wird er diese Information auch nicht an den Versicherer weitergeben. Außerdem sollte das Versicherungsunternehmen stets beachten, dass bei der Erhebung und Verwendung personenbezogener Daten auch Herausforderungen aufgrund des Datenschutzes existieren.

Insgesamt lässt sich also sagen, dass eine statistische Analyse von sozioökonomischen Daten und Krankheitsmerkmalen ein hohes Potenzial bietet, von dem auch die Versicherungswirtschaft profitieren kann. Insbesondere vor dem Hintergrund der zunehmenden Digitalisierung, kann das Sammeln und Auswerten individueller sozioökonomischer Daten in Zukunft ein starkes Werkzeug der Gesundheitsökonomik darstellen.

Die analysierten Daten des SOEP-Datensatzes stellen eine solide Grundlage dar, um einen Krankenhausaufenthalt im Folgejahr abzuschätzen. Abschließend ist jedoch anzumerken, dass zur Verbesserung des Modells der Zugang zu weiteren Daten notwendig ist. Dabei handelt es sich einerseits um konkrete Informationen zu den jeweiligen Kosten eines Krankenhausaufenthaltes bzw. Arztbesuches, andererseits um weitere personenbezogene Daten, die sich auf die Gesundheit einer Person beziehen.

---

[99] Vgl. Mühlbauer, B. / Kellerhoff, F. / Matusiewicz, D. (2014), S. 286

5. Anhang

	2015			
Schulabschluss	Männer	Frauen	Gesamt	Männer
Ohne Schulabschluss	5,79 %	6,57 %	**6,15 %**	5,49 %
Haupt-/Volksschulabschluss	5,64 %	6,13 %	**5,86 %**	5,61 %
Mittlere Reife	4,34 %	5,14 %	**4,71 %**	4,35 %
Abitur/Fachabitur	2,37 %	3,65 %	**2,96 %**	2,35 %

Abbildung 13: Krankenstand nach Schulabschluss in 2015 und 2016[100]

Wie man dieser Abbildung entnehmen kann, sinkt die Anzahl der Tage mit Arbeitsunfähigkeit mit steigendem Schulabschluss kontinuierlich, sodass Personen ohne Abschluss oder mit Haupt-/Volksschulabschluss einen mehr als doppelt so hohen Krankenstand aufweisen als Abiturienten. Darüber hinaus fällt auf, dass Frauen häufiger krank sind als Männer und dass sich der Krankenstand von 2015 zu 2016 leicht verbessert hat.

	2015			
Ausbildungsabschluss	Männer	Frauen	Gesamt	Männer
Ohne beruflichen Ausbildungsabschluss	5,23 %	6,14 %	**5,65 %**	5,14 %
Abschluss einer anerkannten Berufsausbildung	4,62 %	5,02 %	**4,81 %**	4,61 %
Meister-/Techniker- oder gleichwertiger Fachschulabschluss	3,34 %	4,35 %	**3,81 %**	3,31 %
Bachelor	2,33 %	3,53 %	**2,88 %**	2,27 %
Diplom/Magister/Master/ Staatsexamen	1,97 %	3,27 %	**2,57 %**	1,92 %
Promotion	1,45 %	2,37 %	**1,88 %**	1,38 %

Abbildung 14: Krankenstand nach Ausbildungsabschluss in 2015 und 2016[101]

Ähnlich wie Abb. 15, zeigt diese Abbildung, dass der Krankenstand umso besser ist, je höher der berufliche Abschluss ist. Demnach sind diejenigen Personen, die keinen Ausbildungsabschluss haben, deutlich häufiger arbeitsunfähig als Akademiker, insbesondere Promovierte.

[100] Techniker Krankenkasse (2017), S. 70
[101] Techniker Krankenkasse (2017), S. 70

Krankheitskosten
Unter 15 Jahre
15 Jahre bis unter 30 Jahre
30 Jahre bis unter 45 Jahre
45 Jahre bis unter 65 Jahre

Abbildung 15: Krankheitskosten 2015 in EUR in Deutschland nach Alter[102]

In dieser Abbildung sind sowohl Männer als auch Frauen berücksichtigt. Sie zeigt das starke altersbedingte Wachstum der Krankheitskosten. Da der Anteil der älteren Bevölkerung aufgrund des demografischen Wandels ansteigen wird, wird ebenfalls der Mittelwert je Einwohner von derzeit rund 4.140 Euro (Stand: 2015) ansteigen.

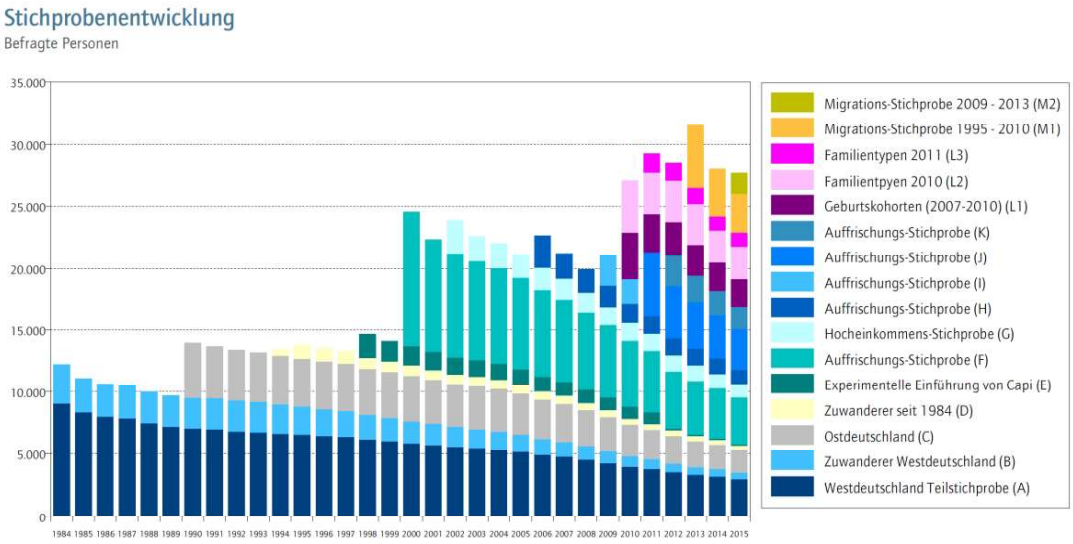

Abbildung 16: Stichprobenentwicklung des Sozio-oekonomischen Panels[103]

Wie man dieser Abbildung entnehmen kann, wird bei der Wahl der Stichproben darauf geachtet, dass insgesamt ein möglichst realistischer Querschnitt der Bevölkerung getroffen wird. Dafür wird stets die aktuelle Situation, auch politisch, berücksichtigt. Außerdem lässt sich deutlich erkennen, wie die einzelnen Stichproben mit der Zeit schrumpfen, was zwangsläufig zusätzliche Stichproben erfordert, damit die Anzahl befragter Personen nicht zu niedrig ist.

[102] Gesundheitsberichterstattung des Bundes (2017)
[103] Deutsches Institut für Wirtschaftsforschung (d)

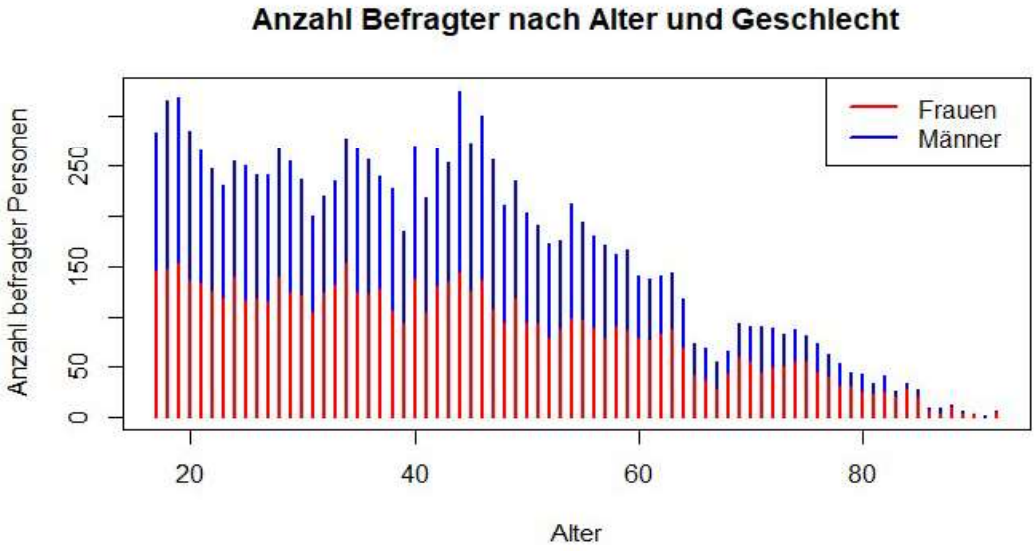

Abbildung 17: Anzahl Befragter nach Alter und Geschlecht, 1984

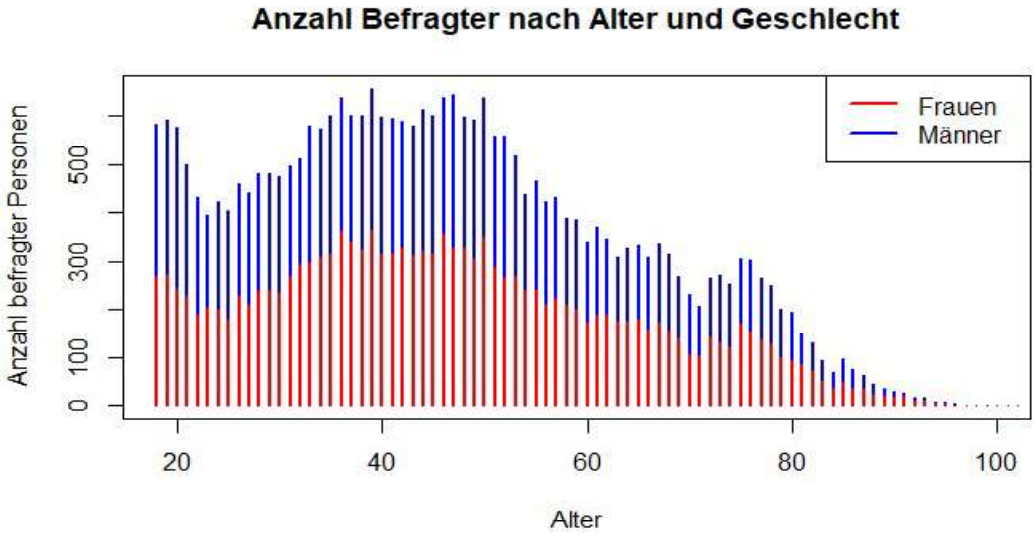

Abbildung 18: Anzahl Befragter nach Alter und Geschlecht, 2016

Vergleicht man die Verteilung der Befragten der Jahre 1984 (Abb. 19) und 2016 (Abb. 20) miteinander, so fällt zunächst auf, dass die Gesamtanzahl der befragten Personen in 2016 deutlich höher war (siehe hierzu auch Abb. 18: Stichprobenentwicklung des Sozio-oekonomischen Panels). Dadurch ist zwar die Volatilität im Jahr 1984 größer, da die Balken der Abbildungen jedoch auch relative Häufigkeiten erkennen lassen, ist eine Vergleichbarkeit der beiden Jahre dennoch möglich. Die Verteilung von 1984 hat noch eher eine „Pyramidenform“ als eine „Urnenform“. Darüber hinaus erkennt man, dass die Bevölkerung im Jahr 2016 insgesamt älter als vor 32 Jahren ist.

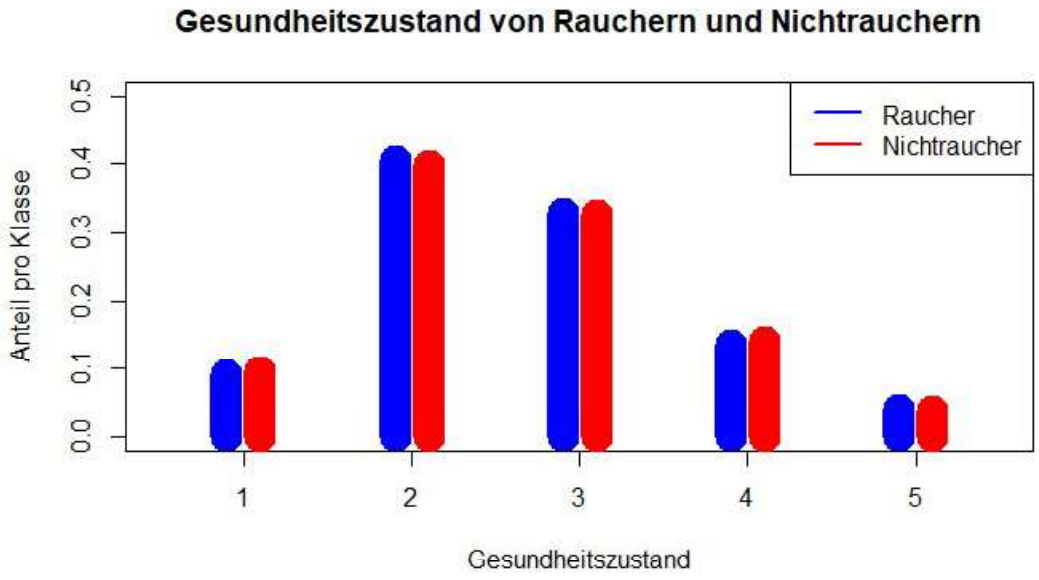

Abbildung 19: Gesundheitszustand von Rauchern und Nichtrauchern

Anders als beim Vergleich zwischen Sportlern und Nichtsportlern, lassen sich in diesem Fall keinerlei merkliche Unterschiede zwischen der Gesundheit von Rauchern und nicht Nichtrauchern feststellen. Sowohl in der ersten Kategorie, dessen Ausprägung „sehr gut“ lautet, als auch in der letzten Kategorie, welche den Gesundheitszustand als „schlecht“ beschreibt, ist der Anteil von Rauchern und Nichtrauchern nahezu identisch.

Alter	Durchschnitts-BMI	
	2002	2016
20	22,05	23,64
40	25,23	26,07
60	26,31	26,88
80	26,2	26,86
Gesamt	**25,29**	**26,35**

Tabelle 11: Durchschnitts-BMI nach Alter, 2002 und 2016

Beim Betrachten der Tabelle ist ein deutlicher Anstieg des BMI sowohl in Bezug auf das steigende Alter als auch bezüglich der Jahre 2002 und 2016 zu erkennen. Entsprechend ergibt sich über alle Altersgruppen für das Jahr 2002 ein Wert von rund 25,29 und für das Jahr 2016, also nur 14 Jahre später, ein mittlerer BMI von 26,35.

In beiden Jahren war demnach der Durchschnitts-BMI oberhalb der Grenze zum Übergewicht.

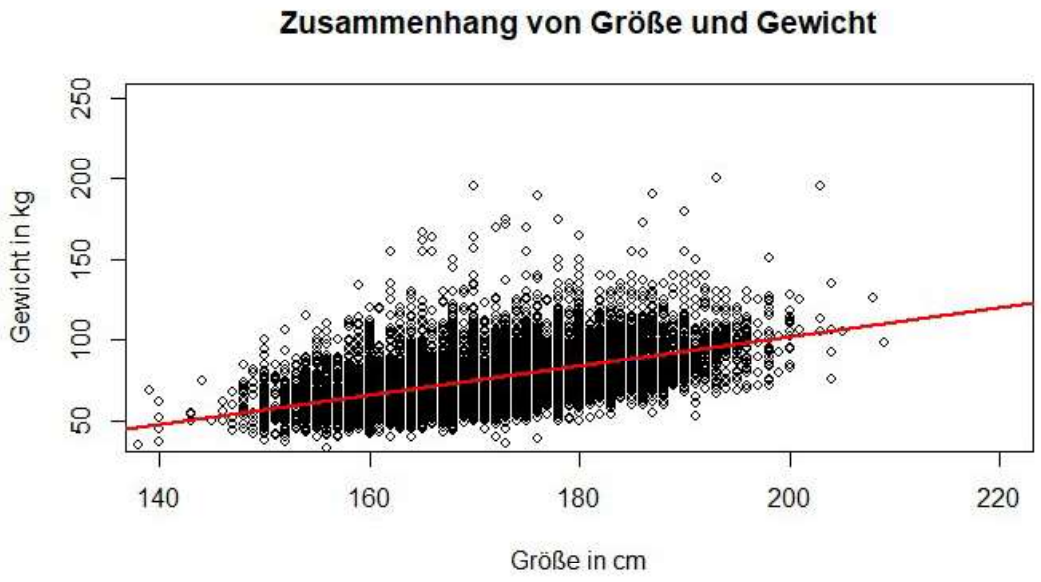

Abbildung 20: Zusammenhang von Größe und Gewicht, LinReg

Anders als in Abbildung 8, liegt bei dieser einfachen linearen Regression eine typische Punktewolke vor, da die Ausprägung des Merkmals Gewicht metrisch skaliert ist. Außerdem sind hier deutlich einzelne Ausreißer zu erkennen. Anhand dieser Wolke kann man bereits erahnen, wie Größe und Gewicht korreliert sind und wie die Steigung der Regressionsgeraden aussehen wird. In diesem Beispiel stellt sich also der Zusammenhang von Größe und Gewicht in der Form dar, dass größere Personen eher mehr wiegen als kleinere.

Ausführliche Herleitung der logistischen Funktion:

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k * x_{ik} + \varepsilon_i = z_i$$

$$\Rightarrow \frac{P(Y=1)}{1-P(Y=1)} = e^{z_i}$$

$$\Rightarrow P(Y=1) = e^{z_i} * (1 - P(Y=1) = e^{z_i} - e^{z_i} * P(Y=1)$$

$$\Rightarrow P(Y=1) + e^{z_i} * P(Y=1) = e^{z_i}$$

$$\Rightarrow P(Y=1) * (1 + e^{z_i}) = e^{z_i}$$

$$\Rightarrow P(Y=1) = \frac{e^{z_i}}{1+e^{z_i}} = \left(\frac{1+e^{z_i}}{e^{z_i}}\right)^{-1} = \left(\frac{1}{e^{z_i}} + \frac{e^{z_i}}{e^{z_i}}\right)^{-1} = \left(1 + \frac{1}{e^{z_i}}\right)^{-1} = \frac{1}{1+e^{-z_i}}$$

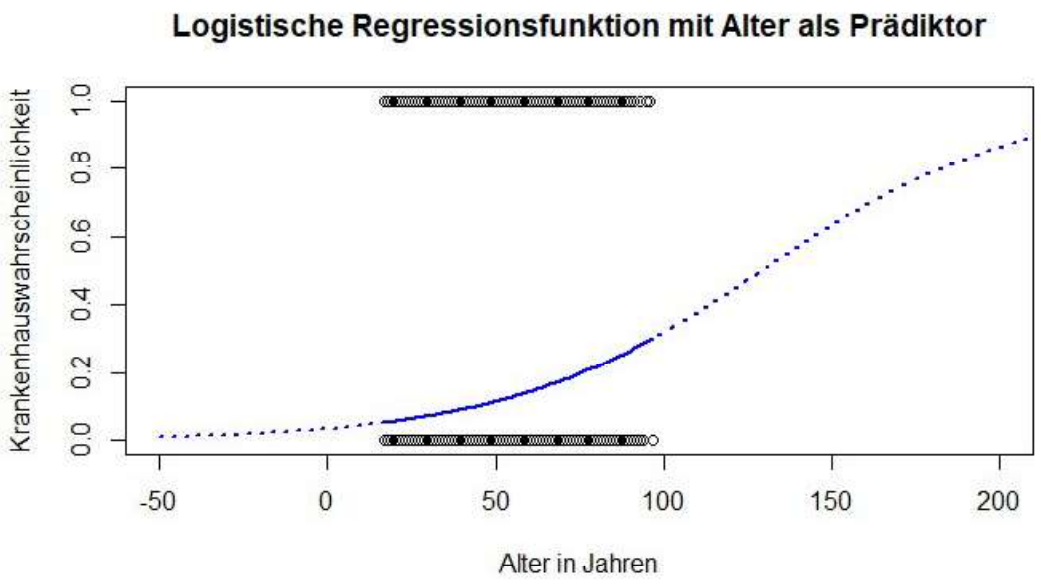

Abbildung 21: Logistische Regressionsfunktion mit Alter als Prädiktor

Zwar sind erreichte Alter jenseits von 120 Jahren extrem selten und unter 0 sogar unmöglich. Dennoch verdeutlicht diese Abbildung gut, wie der typische Verlauf einer logistischen Regressionsfunktion aussieht. Man erkennt, dass sich die Funktionswerte für immer kleiner werdende X-Werte dem Wert 0 und für immer größer werdende X-Werte einer hundertprozentigen Wahrscheinlichkeit annähern.

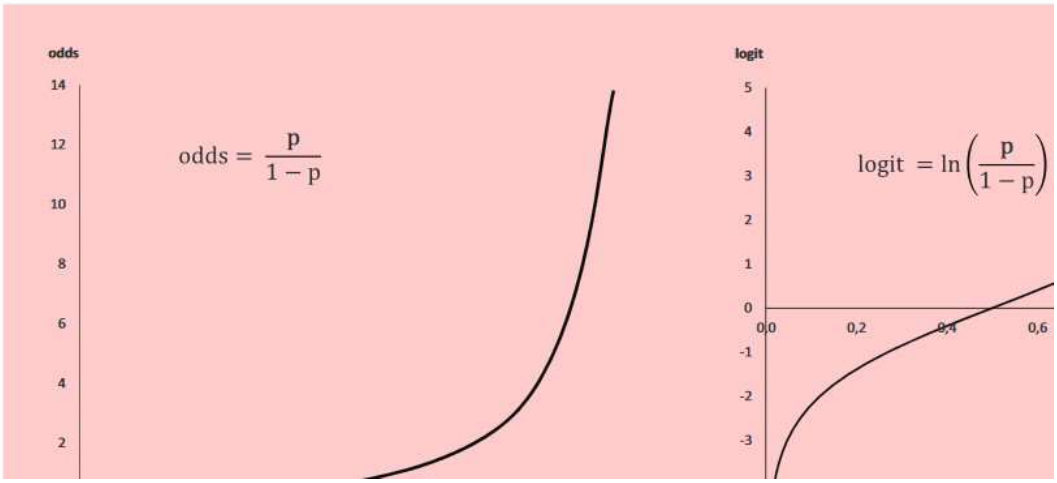

Abbildung 22: Zusammenhang von Odds und Logits[104]

Aus dieser Abbildung geht einerseits hervor, dass die Odds bei steigender Wahrscheinlichkeit sehr stark ansteigen, was darauf zurückzuführen ist, dass durch den Anstieg von $P(Y = 1)$ sowohl der Zähler größer als auch der Nenner kleiner wird. Andererseits lässt sich der Effekt des Logarithmierens der Odds erkennen. Die Logits nehmen negative Werte an, falls die Wahrscheinlichkeiten kleiner als 0,5 sind.

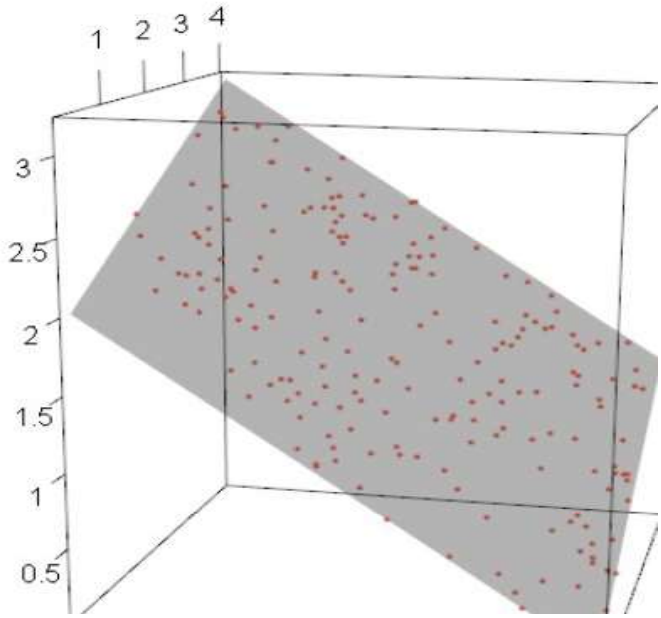

Abbildung 23: Beispiel einer 2-dimensionalen multiplen Regression[105]

[104] Backhaus, K. et al. (2016), S. 293
[105] Wagner, S. (2014)

Wie sich dieser Abbildung entnehmen lässt, kann man multiple lineare Regressionen mit nur zwei Prädiktoren als Ebene in einem dreidimensionalen Koordinatensystem darstellen. Die Übersichtlichkeit ist hier bereits geringer als bei einer einfachen linearen Regressionsgeraden. Nimmt man weitere Prädiktoren hinzu, so wird die Regressionsgleichung zunehmend komplexer und eine übersichtliche grafische Darstellung praktisch nicht mehr möglich.

Merkmal	**Ausprägung**
Geschlecht	Männlich/Weiblich
Job	Ja/Nein
Jobbefristung	Unbefristet/Befristet/kein Arbeitsvertrag/Selbstständig
Monatsgehalt	0 bis 100.000 EUR
Groesse	120 bis 209 cm
Gewicht	33 bis 200 kg
Gesundheitszustand	Sehr gut/Gut/Zufriedenstellend/Weniger gut/Schlecht
Alltagseinschraenkung	Stark eingeschränkt/Etwas eingeschränkt/Nicht eingeschränkt
Geburtsjahr	1909 bis 1989
Diabetes	Ja/Nein
Asthma	Ja/Nein
Herzerkrankung	Ja/Nein
Krebs	Ja/Nein
Schlaganfall	Ja/Nein
Migraene	Ja/Nein
Hypertonie	Ja/Nein
Depression	Ja/Nein
Demenz	Ja/Nein
Gelenkerkrankung	Ja/Nein
Rueckenschmerzen	Ja/Nein
Sonstige_Krankheit	Ja/Nein
Keine_Krankheit	Ja/Nein
KH	Ja/Nein
Anzahl_KH	0 bis 20 Krankenhausaufenthalte
Anzahl_Naechte	0 bis 280 Tage
Raucher	Ja/Nein
Raucher_Startalter	*Im Zeitraum nicht vorhanden*
Bier	Regelmäßig/Ab und zu/Selten/Nie
Wein	Regelmäßig/Ab und zu/Selten/Nie
Spirituosen	Regelmäßig/Ab und zu/Selten/Nie
Ernaehrung	Sehr stark/stark/Ein wenig/Gar nicht
PKV_GKV	GKV/PKV/Weder noch
ZusatzPKV	Ja/Nein
Gesundheitsbedenken	Große Sorgen/Einige Sorgen/Keine Sorgen
Sport	Jede Woche/Jeden Monat/Selten/Nie

Tabelle 12: Gesamtheit aller potenzieller Prädiktoren

Diese Tabelle zeigt alle Merkmale, die als potenzielle Prädiktoren selektiert worden waren, bevor sie teilweise von der statistischen Analyse ausgeschlossen wurden. Teilweise fließen sie in das Basismodell ein, teilweise in das Alternativmodell. Die folgenden Merkmale finden weder beim Basis- noch beim Alternativmodell Anwendung: „Job“, „Jobbefristung“, „Monatsgehalt“, „Alltagseinschränkung“, „Asthma“, „Migräne“, „Demenz“, „Gelenkerkrankung“, „Rückenschmerzen“, „Raucher_Startalter“, „Ernaehrung“, „PKV_GKV“, „ZusatzPKV“. Dies ist in der Regel auf mangelnde Signifikanz dieser Merkmale zurückzuführen. Einige Merkmale hingegen wurden nicht im betrachteten Zeitraum von 2006 bis 2010 abgefragt, wie beispielsweise „Job“, „Alltagseinschränkung“, „Gelenkerkrankung“ und „Raucher_Startalter“.

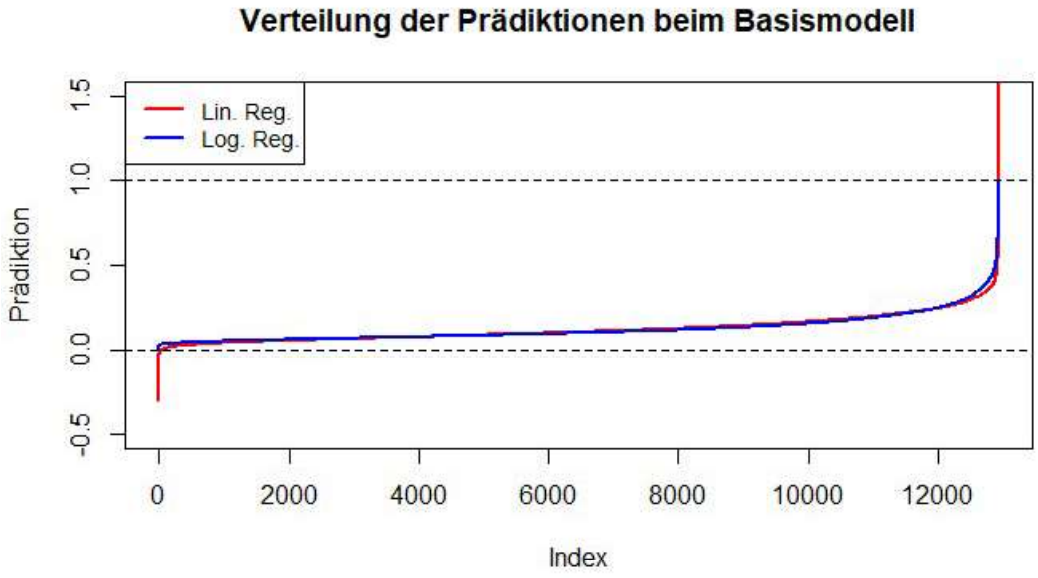

Abbildung 24: Verteilung der Prädiktionen beim Basismodell, 2010

Diese Abbildung verdeutlicht den Nachteil, den die lineare Regression gegenüber der logistischen Regression hat. Sowohl nach unten als auch nach oben gibt es einen Ausreißer, der einen Wert außerhalb des für eine Wahrscheinlichkeit möglichen Bereichs annimmt. Die blaue Kurve hingegen, welche die vorausgesagten Wahrscheinlichkeiten auf Basis einer logistischen Regression darstellt, bewegt sich lediglich im Wertebereich von 0 und 1.

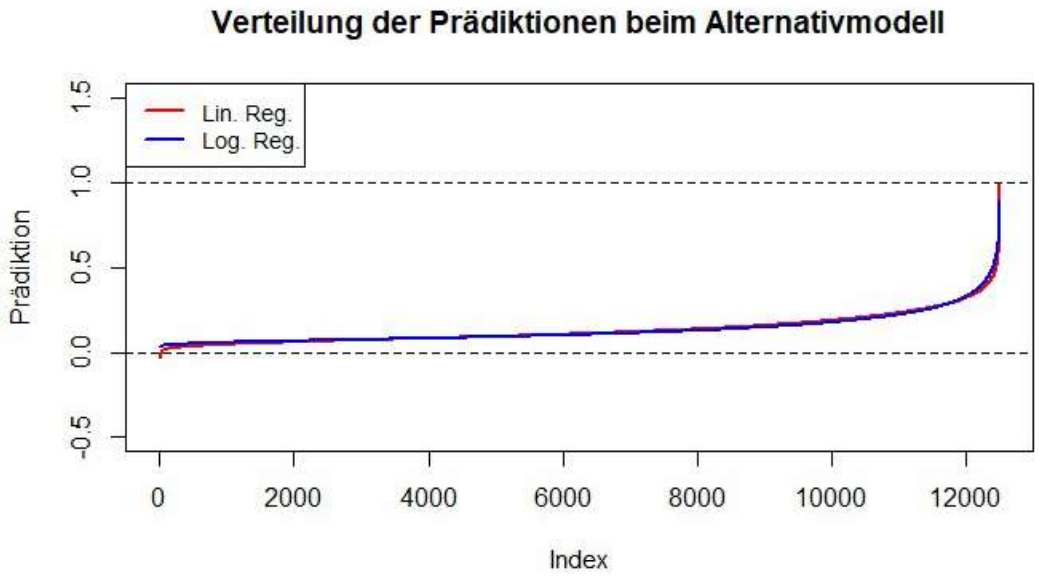

Abbildung 25: Verteilung der Prädiktionen beim Alternativmodell, 2010

Vergleicht man die Verläufe der Kurven des Basismodells (Abb. 26) und die des Alternativmodells (Abb. 27), so kann man bei genauem Betrachten feststellen, dass der Verlauf beim Alternativmodell etwas flacher ist. Entsprechend erreichen die Prädiktionen beim Alternativmodell keine Wahrscheinlichkeiten von fast 100%, sondern nur rund 90%. Außerdem lässt sich auch bei dieser Abbildung erkennen, dass aus der linearen Regression Werte unter 0 bzw. über 1 resultieren können, obgleich diese Ausprägungen nicht so stark sind wie in Abb. 26.

Literaturverzeichnis

Backhaus, Klaus / Erichson, Bernd / Plinke, Wulff / Weiber, Rolf (2016): Multivariate Analysemethoden – Eine anwendungsorientierte Einführung, Springer-Verlag Berlin Heidelberg

Becker, Torsten (2017), Mathematik der private Krankenversicherung, Springer Fachmedien Wiesbaden GmbH, Wiesbaden

Bolte, Gabriele / Kohlhuber, Martina (2008): Untersuchungen der Beiträge von Umweltpolitik sowie ökologischer Modernisierung zur Verbesserung der Lebensqualität in Deutschland und Weiterentwicklung des Konzeptes der Ökologischen Gerechtigkeit - Teilprojekt A: Systematische Zusammenstellung der Datenlage in Deutschland, https://www.umweltbundesamt.de/sites/default/files/medien/publikation/long/3663.pdf

Bös, Klaus / Woll, Alexander (2017): Körperlich Aktive deutlich fitter, Institut für Sport & Sportwissenschaft, Karlsruher Institut für Technologie, https://www.kit.edu/downloads/GzM_Factsheet_Fitness.pdf

Bundesgesundheitsministerium (2018), Gesetzliche Krankenversicherung - Kennzahlen und Faustformeln, https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Statistiken/GKV/Kennzahlen_Daten/KF2018Bund_Juni-2018.pdf

Bundeszentrale für politische Bildung (2018): Bevölkerung mit Migrationshintergrund, http://www.bpb.de/nachschlagen/zahlen-und-fakten/soziale-situation-in-deutschland/61646/migrationshintergrund-i, Zugriff am 20.06.2018

Centor, Robert M. / Schwartz, J. Sanford (1985): An Evaluation of Methods for Estimating the Area Under the Receiver Operating Characteristic (ROC) Curve, in: Med Decis Making, Vol. 5, No. 2, http://journals.sagepub.com/doi/pdf/10.1177/0272989X8500500204

Deutsche Hauptstelle für Suchtfragen e.V. (2018): Pressemitteilung: DHS Jahrbuch Sucht 2018, http://www.dhs.de/fileadmin/user_upload/pdf/news/2018_PM_Daten_und_Fakten.pdf

Deutsches Institut für Wirtschaftsforschung (2015): SOEP-Fragebogen 2016, https://www.diw.de/sixcms/detail.php?id=diw_01.c.499396.de, Zugriff am 01.07.2018

Deutsches Institut für Wirtschaftsforschung (ohne Datum, (a)): Über uns, https://www.diw.de/de/diw_01.c.100293.de/ueber_uns/ueber_uns.html, Zugriff am 01.07.2018

Deutsches Institut für Wirtschaftsforschung (ohne Datum, (b)): Die Survey-Gruppe SOEP, https://www.diw.de/de/diw_02.c.221178.de/ueber_uns.html, Zugriff am 03.07.2018

Deutsches Institut für Wirtschaftsforschung (ohne Datum, (c)): Datenweitergabe 1984-2016 (soep.v33), https://www.diw.de/de/diw_01.c.571790.de/soep_v33.html, Zugriff am 05.07.2018

Deutsches Institut für Wirtschaftsforschung (ohne Datum, (d)): Was ist das Sozio-oekonomische Panel?, https://www.diw.de/sixcms/detail.php?id=299726#299718, Zugriff am 06.07.2018

Deutsches Krebsforschungszentrum (2015): Tabakatlas Deutschland 2015 – Alles auf einen Blick: Zahlen und Fakten, https://www.dkfz.de/de/tabakkontrolle/download/Publikationen/sonstVeroeffentlichungen/Tabakatlas_auf_einen_Blick-Zahlen_und_Fakten.pdf

Dreo, Rainer (2014): Das Working Capital als Indikator für Zahlungsausfälle, Springer Fachmedien, Wiesbaden

Frost, Irasianty (2018): Einfache lineare Regression – Die Grundlage für komplexe Regressionsmodelle verstehen, Springer Fachmedien, Wiesbaden

Gesundheitsberichterstattung des Bundes (2017): Krankheitskosten je Einwohner in € für Deutschland. Gliederungsmerkmale: Jahre, Alter, Geschlecht, ICD10, Einrichtungen, http://www.gbe-bund.de/oowa921-install/servlet/oowa/aw92/WS0100/_XWD_FORMPROC?TARGET=&PAGE=_XWD_102&OPINDEX=1&HANDLER=XS_ROTATE_ADVANCED&DATACUBE=_XWD_130&D.000=ACROSS&D.003=PAGE, Zugriff am 12.08.2018

Hackshaw, Allan / Morris, Joan K. / Boniface, Sadie / Tang, Jin-Ling / Milenković, Dušan (2018): Low cigarette consumption and risk of coronary heart disease and

stroke: meta-analysis of 141 cohort studies in 55 study reports, BMJ 018; 360, https://doi.org/10.1136/bmj.j5855

Handl, Andreas / Kuhlenkasper, Torben (2017): Multivariate Analysemethoden – Theorie und Praxis mit R, 3. Auflage, Springer Spektrum, Berlin

Hartung, Joachim / Elpelt, Bärbel (2007): Multivariate Statistik – Lehr- und Handbuch der angewandten Statistik, 7. Auflage, Oldenbourg Wissenschaftsverlag, München

Kantar Public Deutschland (ohne Datum): Wissen über die Gesellschaft, https://www.tns-infratest.com/sofo/, Zugriff am 29.06.2018

Kotz, Daniel / Böckmann, Melanie / Kastaun, Sabrina (2018): Nutzung von Tabak und E-Zigaretten sowie Methoden zur Tabakentwöhnung in Deutschland, in: Deutsches Ärzteblatt, Jg. 115 Heft 14, https://www.aerzteblatt.de/pdf.asp?id=197190

Kroll, Lars Eric / Lampert, Thomas (2012): Arbeitslosigkeit, prekäre Beschäftigung und Gesundheit, Hrsg. Robert Koch-Institut Berlin, GBE kompakt 3(1), https://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GBEDownloadsK/2012_1_Arbeitslosigkeit_Gesundheit.pdf?__blob=publicationFile

Kumar, Ankit and Michael Schoenstein (2013), "Managing Hospital Volumes: Germany and Experiences from OECD Countries", OECD Health Working Papers, No. 64, OECD Publishing, Paris, https://doi.org/10.1787/5k3xwtg2szzr-en, Zugriff am 17.07.2018

Lampert, Thomas / Kroll, Lars Eric / Müters, Stephan / Stolzenberg, Heribert (2013): Messung des sozioökonomischen Status in der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1), Springer-Verlag, Berlin

Larner, Andrew J. (2015): Optimising the Cutoffs of Cognitive Screening Instruments in Pragmatic Diagnostic Accuracy Studies: Maximising Accuracy or the Youden Index?, in: Dement Geriatr Cogn Disord 2015;39: S. 167–175, https://www.karger.com/Article/Pdf/369883

Mensink, Gert B.M. / Schienkiewitz, Anja / Haftenberger, Marjolein / Lampert, Thomas / Ziese, Thomas / Scheidt-Nave, Christa (2013): Übergewicht und Adipositas in Deutschland - Ergebnisse der Studie zur Gesundheit Erwachsener in Deutschland

(DEGS1),
http://www.gbe-bund.de/pdf/DEGS1_Uebergewicht_Adipositas.pdf

Milbrodt, Hartmut / Kniep, Tobias (2005): Aktuarielle Methoden der deutschen Privaten Krankenversicherung, Verlag Versicherungswirtschaft GmbH, Karlsruhe

Mühlbauer, Bernd H. / Kellerhoff, Fabian. / Matusiewicz, David. (2014): Zukunftsperspektiven der Gesundheitswirtschaft, 2. Auf., LIT Verlag, Berlin

Pfliger, Verena (2014): Bestimmtheitsmaß R^2 - Teil 2: Was ist das eigentlich, ein R^2?, INWT Statistics,
https://www.inwt-statistics.de/blog-artikel-lesen/Bestimmtheitsmass_R2-Teil2.html, Zugriff am 05.08.2018

Robert Koch-Institut (2003): Arbeitslosigkeit und Gesundheit, in: Gesundheitsberichterstattung des Bundes,
Heft 13, http://www.gbe-bund.de/pdf/Heft13.pdf

Robert Koch-Institut (2007): Schimmelpilzbelastung in Innenräumen – Befunderhebung, gesundheitliche Bewertung und Maßnahmen - Mitteilung der Kommission „Methoden und Qualitätssicherung in der Umweltmedizin“, Springer Medizin Verlag, Berlin

Robert Koch-Institut (2015):
https://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GesInDtld/gesundheit_in_deutschland_2015.pdf?__blob=publicationFile

Robin, Xavier / Turck, Natacha / Hainard, Alexandre / Tiberti, Natalia / Lisacek, Frédérique / Sanchez, Jean-Charles / Müller, Markus (2011): pROC: an open-source package for R and S+ to analyze and compare ROC curves,
https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-12-77

Saß, Anke-Christine / Wurm, Susanne / Ziese, Thomas (2005): Alter = Krankheit? Gesundheitszustand und Gesundheitsentwicklung, in: Böhm, Karin / Tesch-Römer, Clemens / Ziese, Thomas (Hrsg.): Gesundheit und Krankheit im Alter, Robert Koch-Institut, Berlin,
http://www.gbe-bund.de/pdf/Gesundh_Krankh_Alter.pdf

Sawitzki, Günther (2008): Statistical Computing – Einführung in R, StatLab Heidelberg

Schlittgen, Rainer (2013): Regressionsanalysen mir R, Oldenbourg Wissenschaftsverlag, München

Schupp, Jürgen et al. (2017): Long format of the German Socio-Economic Panel Study (SOEP), http://dx.doi.org/10.5684/soep.v33

Statistisches Bundesamt (2017): https://www.destatis.de/DE/Publikationen/Thematisch/Gesundheit/Krankenhaeuser/KostennachweisKrankenhaeuser2120630167004.pdf?__blob=publicationFile

Statistisches Bundesamt (2017a): https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2017/11/PD17_399_231.html#Fussnote2, Zugriff am 13.06.2018

Statistisches Bundesamt (2017b): https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/Krankenhaeuser/Tabellen/Diagnosen.html, Zugriff am 18.06.2018

Statistisches Bundesamt (2017c): Statistisches Jahrbuch 2017 – Deutschland und Internationales

Statistisches Bundesamt (2017d): Rauchgewohnheiten nach Altersgruppen und Geschlecht, https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/GesundheitszustandRelevantesVerhalten/Tabellen/Rauchverhalten.html, Zugriff am 23.06.2018

Statistisches Bundesamt (2017e): Verdienste auf einen Blick, https://www.destatis.de/DE/Publikationen/Thematisch/VerdiensteArbeitskosten/Arbeitnehmerverdienste/BroschuereVerdiensteBlick0160013179004.pdf?__blob=publicationFile

Statistisches Bundesamt (2018): https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2018/01/PD18_019_12411.html, Zugriff am 04.07.2018

Techniker Krankenkasse (2017): Gesundheitsreport 2017, https://www.tk.de/centaurus/servlet/contentblob/942842/Datei/63887/Report-AU-Zeiten.pdf

Verband der Privaten Krankenversicherung (2017), Zahlenbericht der Privaten Krankenversicherung 2016,

https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Statistiken/GKV/Kennzahlen_Daten/KF2018Bund_Juni-2018.pdf

Wagner, Sarah (2014): Multiple lineare Regression, INWT Statistics

Wagner, Sarah (2015): Logistische Regression - Modell und Grundlagen, INWT Statistics, https://www.inwt-statistics.de/blog-artikel-lesen/Logistische_Regression.html, Zugriff am 03.08.2018

Wentura, Dirk / Pospeschill, Markus (2015): Multivariate Datenanalyse – Eine kompakte Einführung, Springer Fachmedien, Wiesbaden

World Health Organization (2017): Overweight and Obesity in the Western Pacific Region - An Equity Perspective, http://apps.who.int/iris/bitstream/handle/10665/255475/9789290618133-eng.pdf

Ehrenwörtliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

[redacted], den 31. August 2018